LICENCE EN BIOLOGIE

Caractérisation structurale d'un régulateur transcriptionnel du « quorum sensing » chez Brucella abortus

Wenders, Olivier

Award date:
2000

Awarding institution:
Universite de Namur

Piera PORFIDO-BON
SECRETARIAT BIOLOGIE F.u.N.D.P.
Rue de Bruxelles, 59
B-5000 NAMUR (Belgique)
Tél. 081/72 44 26 - Fax 081/72 44 20

FACULTES UNIVERSITAIRES NOTRE-DAME DE LA PAIX
NAMUR

Faculté des Sciences

Caractérisation structurale d'un régulateur transcriptionnel du « quorum sensing » chez *Brucella abortus*

Mémoire présenté pour l'obtention du grade de licencié en Sciences biologiques

Olivier Wenders

Juin 2000

Facultés Universitaires Notre-Dame de la Paix
FACULTE DES SCIENCES
Secrétariat du Département de Biologie
Rue de Bruxelles 61 - 5000 NAMUR
Téléphone: + 32(0)81.72.44.18 - Téléfax: + 32(0)81.72.44.20
E-mail: joelle.jonet@fundp.ac.be - http://www.fundp.ac.be/fundp.html

Caractérisation structurale d'un régulateur transcriptionnel du « Quorum Sensing » chez *Brucella abortus*.

WENDERS Olivier

Résumé

Le « Quorum Sensing » est un mécanisme régulatoire dépendant de la densité cellulaire, utilisé par de nombreuses bactéries *Gram-négatives*. Une phéromone nommée N-acyl-L-homosérine lactone (HSL ou autoinducteur) est produite par les cellules bactériennes à un taux basal. Quand la densité cellulaire dépasse un certain seuil, cette phéromone atteint alors une concentration suffisante pour activer un régulateur transcriptionnel. Le régulateur de « Quorum Sensing » le plus étudié à ce jour est le régulateur LuxR de *Vibrio fischeri* qui peut être divisé en deux domaines structuraux. Le domaine N-terminal est impliqué dans la liaison à l'autoinducteur et le domaine C-terminal dans la liaison à l'ADN (Fuqua, Winans et al. 1994). Des systèmes de régulation homologues ont été identifiés dans d'autres bactéries *Gram-negatives*. La structure tridimensionnelle de ces régulateurs est inconnue à ce jour.

Dans ce travail, nous avons utilisé des outils bioinformatiques pour prédire des caractéristiques structurales de BabR, un homologue de LuxR récemment identifié chez la bactérie pathogène *Brucella abortus*. Un modèle du domaine C-terminal a pu être construit sur base de l'homologie avec NarL, une protéine d'*E. coli* intervenant dans un système de régulation à deux composants. Une analyse de différents complexes HTH-ADN a permis de récolter des informations quant au domaine HTH du modèle de BabR. Nous avons également obtenu un modèle topologique du domaine N-terminal de BabR. Ceci est intéressant car nous avons pu fournir une hypothèse sur la structure 3D de ce domaine, ce qui n'avait encore jamais été décrit.

Mémoire de licence en Sciences biologiques
Juin 2000
Promoteur: E. Depiereux
Co-Promoteur : J.J. Letesson

A l'issue de ce travail, je tiens tout d'abord à remercier chaleureusement M. le Professeur Depiereux de m'avoir accueilli dans son laboratoire

Je tiens ensuite a remercier tout particulièrement Katalin de m'avoir aidé à surmonter cette épreuve Ô combien passionante

Il y a aussi tout BMS qui m'a fortement aidé et …

Excusez-moi , je suis comme on dirait à « la bourre »

J'aimerais en remercier tellement d'autre comme ma famille mais je suis en retard…

ERRATA

- La référence correspondant au tableau 1 à la page 6 ne doit pas se trouver dans le paragraphe concernant les γ-butyrolactones mais bien dans le dernier paragraphe de la page 6 (oligopeptides)
- Les tableaux 15 et 16 (partie résultats et discussion) ne sont pas corrects : ils doivent être remplacés par les tableaux 15a, 15b et 16a et 16b ci-joints.
- Figure 36 : légende à compléter : Les triangles orientés dans le même sens représentent un plan β parallèle et les cercles, des hélices. Les points noirs correspondent, de l'extrémité N-terminale à l'extrémité C-terminale, aux résidus D76, L79, W91, I102, L111, G114, S116 de BabR.

3 CRO
Monom. 1

Struct. Sec.	A. aminés
	THR
H	GLN
H	THR
H	GLU
H	LEU
H	ALA
H	THR
H	LYS
H	ALA
T	GLY
T	VAL
T	LYS
T	GLN
H	GLN
H	SER
H	ILE
H	GLN
H	LEU
H	ILE
H	GLU
H	ALA
	GLY

3 CRO
Monom. 2

Struct. Sec.	A. aminés
H	GLN
H	THR
H	GLU
H	LEU
H	ALA
H	THR
H	LYS
H	ALA
T	GLY
T	VAL
T	LYS
T	GLN
H	GLN
H	SER
H	ILE
H	GLN
H	LEU
H	ILE
H	GLU
H	ALA
	GLY

1 QAA
Monom. 1

Struct. Sec.	A. aminés
	THR
H	ARG
H	ALA
H	GLU
H	ILE
H	ALA
H	GLN
H	ARG
T	LEU
T	GLY
T	PHE
T	ARG
T	SER
T	PRO
H	ASN
H	ALA
H	ALA
H	GLU
H	GLU
H	HIS
H	LEU
H	LYS
H	ALA
H	LEU
H	ALA
H	ARG
H	LYS
	GLY

1 QAA
Monom. 2

Struct. Sec.	A. aminés
	THR
H	ARG
H	ALA
H	GLU
H	ILE
H	ALA
H	GLN
H	ARG
T	LEU
T	GLY
T	PHE
T	ARG
T	SER
T	PRO
H	ASN
H	ALA
H	ALA
H	GLU
H	GLU
H	HIS
H	LEU
H	LYS
H	ALA
H	LEU
H	ALA
H	ARG
H	LYS
	GLY

1 QAR
Monom. 1

Struct. Sec.	A. aminés
	ARG
H	GLN
H	ALA
H	ALA
H	LEU
H	GLY
H	LYS
H	MET
H	VAL
T	GLY
T	VAL
T	SER
T	ASN
H	VAL
H	ALA
H	ILE
H	SER
H	GLN
H	TRP
H	GLU
H	ARG
	SER
	GLU
	THR

1 QAR
Monom. 2

Struct. Sec.	A. aminés
	ARG
H	GLN
H	ALA
H	ALA
H	LEU
H	GLY
H	LYS
H	MET
H	VAL
T	GLY
T	VAL
T	SER
T	ASN
H	VAL
H	ALA
H	ILE
H	SER
H	GLN
H	TRP
H	GLU
H	ARG
	SER
	GLU
	THR

1 LMD
Monom. 1

Struct. Sec.	A. aminés
	SER
H	GLN
H	GLU
H	SER
H	VAL
H	ALA
H	ASP
H	LYS
H	MET
T	GLY
T	MET
T	GLY
H	GLN
H	SER
H	GLY
H	VAL
H	GLY
H	ALA
H	LEU
H	PHE
	ASN

1 LMD
Monom. 2

Struct. Sec.	A. aminés
	SER
H	GLN
H	GLU
H	SER
H	VAL
H	ALA
H	ASP
H	LYS
H	MET
T	GLY
T	MET
T	GLY
H	GLN
H	SER
H	GLY
H	VAL
H	GLY
H	ALA
H	LEU
H	PHE
	ASN

Tableau 15a : Résidus des 4 complexes analysés, interagissant spécifiquement avec l'ADN

lux R

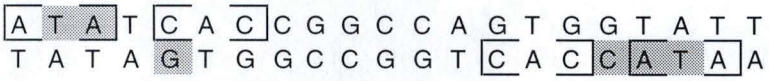
Struct. Sec.	A. aminés
	SER
H	SER
H	TRP
H	ASP
H	ILE
H	SER
H	LYS
H	ILE
H	LEU
T	GLY
T	CYS
T	SER
H	LYS
H	ARG
H	THR
H	VAL
H	THR
H	PHE
H	HIS
H	LEU
H	THR
H	ASN
H	ALA
H	GLN
H	MET
H	LYS
H	LEU
	ASN

Bab R

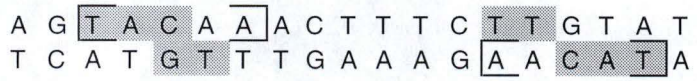
Struct. Sec.	A. aminés
	THR
H	ALA
H	GLU
H	ILE
H	ILE
H	GLY
H	THR
H	ILE
H	LEU
T	ASN
T	ILE
T	SER
H	THR
H	ARG
H	THR
H	VAL
H	ASN
H	PHE
H	HIS
H	ILE
H	ASN
H	ASN
H	VAL
H	LEU
H	THR
H	LYS
H	LEU
	VAL

Tableau 15b : Prédiction des zones du HTH des régulateurs LuxR et BabR interagissant spécifiquement avec l'ADN

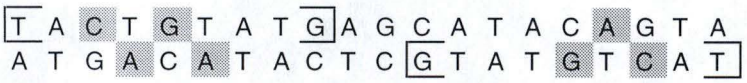
1LMD



3CRO



1QAA



1QAR

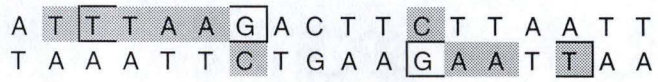


Tableau 16a : X = base interagissant avec le domaine HTH de la protéine du complexe

[ ] = région palindrômique

***lux* box**

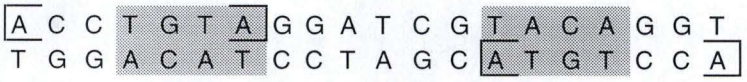


Tableau 16b : Les données de ce tableau ont été prédites sur base des données du Tableau 16a
La région en gris représente la région de la *lux* box prédite comme pouvant interagir avec le HTH. Les centres des deux boîtes sont séparés par 10 résidus et chacune de celles-ci comprend un motif TA

[ ] = région palindrômique

TABLE des MATIERES

Figure 1 : Les points noirs représentent les fortes concentrations bactériennes dans la nature (Sonea and Panisset 1980)

PREFACE

C'est un matin en juin 1676 qu'un homme observa, probablement pour la première fois, des bactéries. Le hollandais Antonie van Leewenhoek, qui eut le bonheur de faire cette expérience, décrit les bactéries comme des « créatures vivantes ressemblant à des grains de sable vus à l'œil nu ». Nous savons aujourd'hui que les bactéries étaient parmi les premiers êtres vivants à habiter notre planète et qu'elles colonisent les milieux les plus variés (figure 1). Leur diversité est phénoménale, nous ne connaissons qu'une infime partie du monde bactérien : la zone émergée de l'iceberg représente sûrement moins de 2% du total.

Leur constitution est très simple et elles ne renferment pas de structures élaborées telles qu'un cytosquelette complexe ou des organites spécialisés comme le noyau ou l'appareil de Golgi. Malgré cela, les bactéries s'adaptent à merveille à leur environnement (Losick and Kaiser, 1997 pour une revue).

Les bactéries ne s'organisent pas en organes ou en tissus, mais elles constituent pourtant un super-organisme vieux de trois milliards d'années. Elles sont, en effet, capables de fonctionner de façon solidaire et de s'échanger des informations. Les mécanismes qui gèrent les fonctions communes ne rappellent en rien ceux des organismes eucaryotes supérieurs dont les cellules sont reliées les unes aux autres par le système nerveux et endocrinien. Dans le cas des bactéries, les liens sont des molécules d'information et de communication. Il s'agit essentiellement de matériel génétique renfermant toute l'information héréditaire. De ce fait, les cellules isolées ont accès, *via* les très nombreux éléments mobiles, à un énorme génome bactérien qui encode plus de fonctions que ne le peut le génome cellulaire individuel (Sonea and Panisset, 1980). Ainsi, bien que les sciences du 20ème siècle, et la microbiologie en particulier, soient caractérisées par une stratégie réductionniste qui a permis de comprendre le fonctionnement détaillé de la cellule, il faut se rendre compte qu'une bactérie n'est pas qu'un organisme unicellulaire mais un membre à part entière d'une communauté microscopique et macroscopique avec toutes les interactions entre ses composants qu'elle implique. Nous ne pouvons aborder ce sujet sans en revenir à Robert Koch, un des pères de la bactériologie.

INTRODUCTION

1. Les bactéries et la multicellularité

1.1. Contexte historique

L'époque *pré-Kochienne* était caractérisée par la domination, en microbiologie, des pléomorphistes menés par le botaniste suisse Carl Von Naegli. Ceux-ci clamaient haut et fort qu'il n'y a pas de distinction d'espèces chez les bactéries et que c'est leur capacité à se transformer en tout autre microbe qui explique la succession des types dans une culture mixte. Mais Robert Koch comprit que cette notion de pléomorphisme était inconsistante et en opposition avec le concept de métabolisme spécifique commençant à émerger des travaux de Louis Pasteur. Cela l'amena aussi à croire que les infections pathologiques ne pourraient être comprises qu'en ayant à l'esprit la notion de spécificité bactérienne et en utilisant comme moyen d'étude la culture pure.

Ces idées monomorphistes s'instaurèrent plus tard en de véritables dogmes et Koch devint un scientifique autoritaire et influent. Malheureusement, il emmena la microbiologie vers une impasse malgré tout ce qu'il lui apporta de bon. En effet, il devint en quelque sorte victime de son insistance à éliminer autant de variables que possible et à ne voir les bactéries que comme des formes stables aux fonctions stables. Ce dogme a rendu difficile la reconnaissance de la réalité des interactions des bactéries entre-elles, avec d'autres bactéries, avec d'autres organismes et avec leur substrat. De plus, la sanctification de la culture pure ne permit pas d'avoir accès à tout ce qui se passe en dehors du laboratoire, dans les populations naturelles souvent mixtes.

Le paradigme de Koch ne resta pas incontesté durant tout ce temps. Le microbiologiste russe Sergei Winogradsky emprunta une voie alternative à celle de Koch pour son étude des populations microbiennes du sol. Pour lui, elles devaient être étudiées comme un tout puisque « la compétition entre ses composants est le déterminant principal de leurs fonctions individuelles ». Il resta ainsi un fervent adepte de la méthode botanique classique d'observation directe de cultures mixtes, bien qu'il n'abandonnât pas la culture pure pour la compréhension des processus métaboliques. Il y en eu d'autres qui, comme Winogradsky, n'acceptèrent pas aveuglément le dogme de Koch comme Henrici, Waksman ou Dubos. Un des

Bacterial species	Signal	Molecular class	Phenotype affected
Staphylococcus aureus	Rap	Octapeptide	Toxic exoprotein, virulence factor secretion
Bacillus subtilis	ComX	Decapeptide, modified tryptophan	Transformation competence
Bacillus subtilis	CSF	Pentapeptide	Transformation competence, sporulation
Bacillus subtilis	Sporulation factor	Oligopeptide	Sporulation
Streptococcus pneumoniae	ComC	Heptadecapeptide	Transformation competence
Lactococcus lactis	Nisin	Oligopeptide	Nisin (lantibiotic) production
Lactobacillus plantarum	Bacteriocin inducing factor	26 amino acid oligopeptide	Plantaricin (class II antimicrobial)
Lactobacillus sake	Bacteriocin inducing factor	Oligopeptide	Sakacin (class II antimicrobial)
Carnobacterium piscicola	Bacteriocin inducing factors	24 and 49 amino acid oligopeptides	Carnobacteriocin (class II antimicrobial)
Enterococcus faecalis	Sex pheromones	Oligopeptides, plasmid-specific	Agglutination, plasmid transfer
Enterococcus faecalis	Sex pheromone inhibitors	Oligopeptides, plasmid-specific	Inhibit sex pheromone binding

Tableau 1 : Différents types d'oligopeptides de *Gram-negatives* et de *Gram-positives* (Shapiro 1998)

INTRODUCTION

1. Les bactéries et la multicellularité

1.1. Contexte historique

L'époque *pré-Kochienne* était caractérisée par la domination, en microbiologie, des pléomorphistes menés par le botaniste suisse Carl Von Naegli. Ceux-ci clamaient haut et fort qu'il n'y a pas de distinction d'espèces chez les bactéries et que c'est leur capacité à se transformer en tout autre microbe qui explique la succession des types dans une culture mixte. Mais Robert Koch comprit que cette notion de pléomorphisme était inconsistante et en opposition avec le concept de métabolisme spécifique commençant à émerger des travaux de Louis Pasteur. Cela l'amena aussi à croire que les infections pathologiques ne pourraient être comprises qu'en ayant à l'esprit la notion de spécificité bactérienne et en utilisant comme moyen d'étude la culture pure.

Ces idées monomorphistes s'instaurèrent plus tard en de véritables dogmes et Koch devint un scientifique autoritaire et influent. Malheureusement, il emmena la microbiologie vers une impasse malgré tout ce qu'il lui apporta de bon. En effet, il devint en quelque sorte victime de son insistance à éliminer autant de variables que possible et à ne voir les bactéries que comme des formes stables aux fonctions stables. Ce dogme a rendu difficile la reconnaissance de la réalité des interactions des bactéries entre-elles, avec d'autres bactéries, avec d'autres organismes et avec leur substrat. De plus, la sanctification de la culture pure ne permit pas d'avoir accès à tout ce qui se passe en dehors du laboratoire, dans les populations naturelles souvent mixtes.

Le paradigme de Koch ne resta pas incontesté durant tout ce temps. Le microbiologiste russe Sergei Winogradsky emprunta une voie alternative à celle de Koch pour son étude des populations microbiennes du sol. Pour lui, elles devaient être étudiées comme un tout puisque « la compétition entre ses composants est le déterminant principal de leurs fonctions individuelles ». Il resta ainsi un fervent adepte de la méthode botanique classique d'observation directe de cultures mixtes, bien qu'il n'abandonnât pas la culture pure pour la compréhension des processus métaboliques. Il y en eu d'autres qui, comme Winogradsky, n'acceptèrent pas aveuglément le dogme de Koch comme Henrici, Waksman ou Dubos. Un des

exemples les plus frappants de l'interaction entre deux espèces de microbes est l'inhibition de l'un par une substance produite par l'autre, telle qu'un antibiotique.

L'intérêt pour les relations microbe-microbe fut relancé grâce aux travaux de Gause suivis par ceux de Jacques Monod qui essayèrent de comprendre ces interactions au moyen d'analyses mathématiques. Bien que n'aboutissant pas à des résultats concluants, ces travaux furent une marche en avant pour les découvertes des années 60, notamment sur les interactions compétitives entre bactéries marines. Ainsi, Powell prédit ces résultats quelques années plus tôt à l'aide de méthodes mathématiques.

Malgré cela, l'étude des relations microbe-microbe resta, jusqu'il y a un peu plus d'une dizaine d'années, l'une des facettes les plus floues de la microbiologie. En effet, l'étude des populations mixtes a pris du temps à se développer et à se faire accepter par les scientifiques (Dworkin, 1997).

Depuis, nos connaissances sur le sujet ont fortement évolué et les communications intercellulaires sont aujourd'hui considérées comme un trait général chez les bactéries. Ainsi, de nombreuses classes de molécules de communication ont été identifiées chez les espèces *Gram-positives* et *Gram-négatives*. Ces molécules, grâce à des réseaux complexes de transduction du signal, contrôlent l'expression de gènes intervenant dans des processus comme l'interaction avec l'hôte ou la différenciation cellulaire.

Ce bouleversement des idées en microbiologie est principalement dû à la découverte de molécules impliquées dans un phénomène dépendant de la densité (voir chapitre suivant). En particulier les phéromones des espèces *Gram-négatives*, utilisées d'un bout à l'autre du royaume eubactérien pour réguler l'expression d'une large variété de phénotype (Dunlap, 1997),(Fuqua *et al.*, 1994). La dernière décennie fut aussi le témoin de la découverte de phénomènes tels que l'autoagrégation de bactéries chémotactiques (Budrene and Berg, 1995) ou les comportements coordonnés dans la morphogenèse de colonies complexes (Ben-Jacob and Cohen, 1997) ou encore de la prédominance des biofilms dans tous les écosystèmes (Costerton *et al.*, 1995). De plus, les bases moléculaires des communications intercellulaires ont été clarifiées et plusieurs avantages ont été associés aux coopérations multicellulaires :

- Une prolifération plus efficace grâce à une coopération pour la division cellulaire
- Un accès à des ressources et à des niches qui ne peuvent être accessibles à des cellules isolées
- Une défense collective contre des agresseurs que ne peuvent soutenir des cellules isolées
- Une optimalisation de la survie de la population par une différenciation en types cellulaires distincts

Bacterial species	Signal	Molecular class	Phenotype affected
Staphylococcus aureus	Rap	Octapeptide	Toxic exoprotein, virulence factor secretion
Bacillus subtilis	ComX	Decapeptide, modified tryptophan	Transformation competence
Bacillus subtilis	CSF	Pentapeptide	Transformation competence, sporulation
Bacillus subtilis	Sporulation factor	Oligopeptide	Sporulation
Streptococcus pneumoniae	ComC	Heptadecapeptide	Transformation competence
Lactococcus lactis	Nisin	Oligopeptide	Nisin (lantibiotic) production
Lactobacillus plantarum	Bacteriocin inducing factor	26 amino acid oligopeptide	Plantaricin (class II antimicrobial)
Lactobacillus sake	Bacteriocin inducing factor	Oligopeptide	Sakacin (class II antimicrobial)
Carnobacterium piscicola	Bacteriocin inducing factors	24 and 49 amino acid oligopeptides	Carnobacteriocin (class II antimicrobial)
Enterococcus faecalis	Sex pheromones	Oligopeptides, plasmid-specific	Agglutination, plasmid transfer
Enterococcus faecalis	Sex pheromone inhibitors	Oligopeptides, plasmid-specific	Inhibit sex pheromone binding

Tableau 1 : Différents types d'oligopeptides de *Gram-negatives* et de *Gram-positives* (Shapiro 1998)

Les études sur les comportements collectifs et sur les molécules de signal se poursuivent avec intérêt. Elles nous permettent d'entrevoir l'importance des communications bactériennes dans la biosphère. La multi-cellularité repose sur la capacité des bactéries individuelles à recevoir, à interpréter et à répondre aux informations provenant de leurs voisins (Shapiro, 1998).

1.2. Diverses classes de molécules de signal

Les N-acyl-L-homosérines lactones des bactéries *Gram-négatives* ont une fonction d'activateurs de régulateurs de la transcription. Ces phéromones interviennent dans un phénomène dépendant de la densité que nous avons évoqué dans le paragraphe précédent et qui se nomme le Quorum Sensing (Fuqua *et al.*, 1994). Celui-ci permet à la cellule individuelle d'être assurée d'une densité de population suffisante avant qu'elle ne lance sa machinerie d'expression spécifique à certaines fonctions (voir chapitre suivant). La plupart des N-acyl-L-homosérines lactones (AHLs) stimulent leur propre production, d'où leur nom « d'autoinducteur ». Ces signaux sont, en général, des facteurs diffusibles (Dunlap, 1997),(Fuqua *et al.*, 1994).

Les γ-butyrolactones des espèces du genre *Streptomyces* (tableau1) sont considérées comme des analogues des AHLs (Beppu, 1995) mais d'autres molécules en sont distinctes, incluant celles qui sont communément classées parmi les toxines (les antibiotiques et les bactériocines).

A l'heure actuelle, aucune AHL synthétisée par bactéries *Gram-positives* n'a encore été trouvée. Les molécules de signal les plus largement utilisées parmi les bactéries *Gram-positives* sont des oligopeptides stimulant souvent aussi leur propre production. La grande diversité des oligopeptides les rend particulièrement bien adaptés quand une forte discrimination entre molécules est requise.

> Exemple : C'est le cas de la régulation du transfert de plasmides conjugatifs entre des coques *Gram-positives* impliquant différents oligopeptides d'accouplement (Clewell, 1993).

A la différence de la plupart des AHLs, les oligopeptides ne traversent pas librement les membranes bactériennes mais sont exportés de la cellule par des ABC transporteurs (ATP-Binding Cassette transporteurs) et sont détectés par des récepteurs appartenant au système de régulation à deux composantes[1] (Kleerebezem *et al.*, 1997). Ces oligopeptides sont encodés par des gènes, alors que les AHLs sont produites par une enzyme. Ce système de communication basé sur les peptides se

[1] C'est un mécanisme de transduction du signal, retrouvé dans plus de 60 systèmes régulatoires bactériens. Il fait intervenir un senseur qui est une histidine kinase autophosphorylante et un régulateur qui est activé par la phosphorylation.

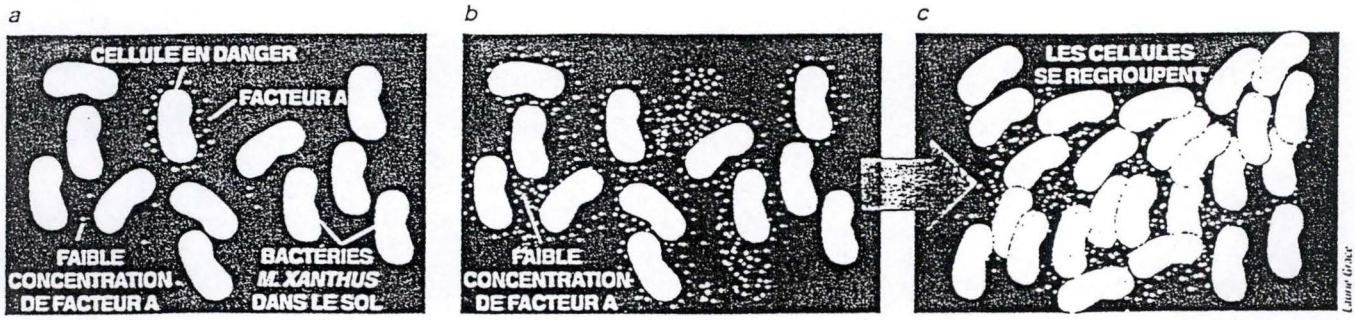


Figure 2 : (Losick and Kaiser 1997)

LES BACTÉRIES DU SOL de l'espèce *Myxococcus xanthus* vivent en petits groupes d'individus tant que les ressources alimentaires sont suffisantes. Lorsque celles-ci s'amenuisent et que les cellules commencent à souffrir d'inanition, elles signalent leur détresse en sécrétant une substance chimique nommée facteur A. Une faible concentration en facteur A *(a)* n'a que peu d'effets sur une population bactérienne, mais lorsque cette concentration dépasse une valeur seuil *(b)*, cette concentration indique à la population qu'elle est en danger et incite les cellules bactériennes à se regrouper *(c)* pour former un corps fructifère . De tels corps fructifères contiennent jusqu'à 100 000 spores, qui résistent à un manque de nourriture. Ces spores sont alors transportées en bloc par le vent ou par un animal vers des zones plus favorables, où elles peuvent germer et établir une nouvelle colonie.

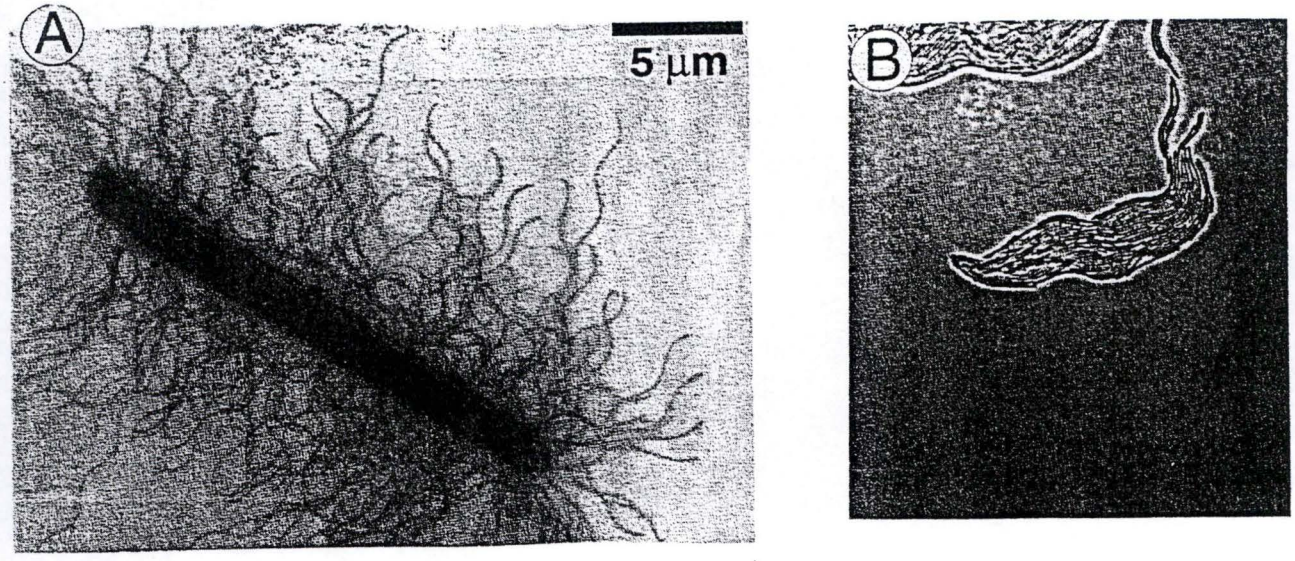


Figure 3 : Le swarming chez *Proteus mirabilis* (A) Cellule swarmer complètement différenciée vue par microscopie électronique (B) Mouvement de masse de cellules swarmers sur agar (Belas 1997)

retrouve aussi chez certaines *Gram-négatives* (Dunny and Winans, 1999). D'ailleurs, on a montré qu'il existe une homologie entre certains récepteurs des *Gram-positives* et ceux de bactéries entériques (*Gram-négatives*).

Certains signaux peptidiques sont diffusibles. Ainsi, ce sont des peptides diffusibles qui servent de molécules de communication dans la formation de corps fructifères chez *Myxococcus xanthus* (Kaplan and Plamann, 1996) et durant l'autoagrégation chez *Escherichia coli* (Budrene and Berg, 1995). Par exemple, le facteur A de *Myxococcus xanthus* active un système de transduction du signal quand la densité cellulaire dépasse un certain seuil, ce qui aboutit à l'expression de fonctions nécessaires au développement du corps fructifère (Kaplan and Plamann, 1996) (figure 2).

1.3. Les comportements multicellulaires : quelques exemples

1.1.1. Le swarming

Le swarming est un processus observé aussi bien chez des espèces *Gram-négatives* que chez des espèces *Gram-positives* et qui consiste en la migration rapide sur une surface d'un groupe de cellules « swarmers » allongées, hyperflagellées et encapsulées dans des exopolymères. Ces cellules sont pourvues de plusieurs nucléoïdes. Des bactéries comme *Proteus vulgaris*, *Proteus mirabilis*, *Serratia marcescens* illustrent bien le swarming (Belas, 1997) (figure 3).

Leur longueur et leur hyperflagellation les distinguent des cellules « swimmers » qui ne savent se déplacer que dans un milieu fluide et pas sur une surface d'agar. Ce phénomène de swarming ne se produit que pour un groupe de cellules. Une cellule isolée ne se déplacera pas. La capsule polysaccharidique que produit *Proteus mirabilis* joue un rôle majeur dans le swarming car des mutants ne sécrétant pas ces polymères ne migreront pas sur l'agar (Gygi *et al.*, 1995). Chez *Serratia liquefaciens*, il a été montré que le swarming est contrôlé, entre autres, par un signal de type AHL (Givskov *et al.*, 1998).

1.1.2.L'utilisation de ressources et l'accès aux niches écologiques

Les polymères organiques complexes doivent être dégradés avant de pouvoir être utilisés par des bactéries. Beaucoup d'entre-elles requièrent une action concertée des individus de la population pour cette dégradation. Ainsi, les myxobactéries telles que *Myxococcus xanthus* lysent leurs proies en libérant des enzymes digestives extracellulaires. Mais l'environnement aqueux peut diluer ces exoenzymes et donc, pour assurer une action efficace, elles forment des colonies sphériques dans lesquelles elles emprisonnent les organismes et les lysent. Ce comportement

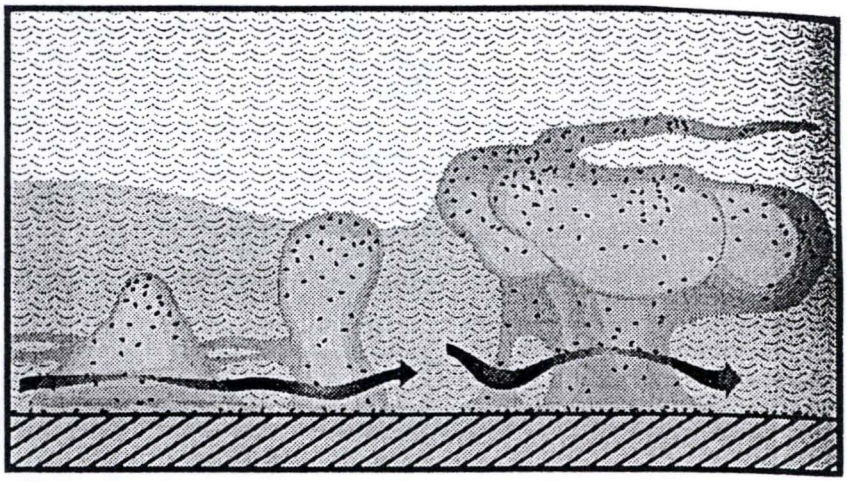


Figure 4 : Modèle conceptuel de l'architecture d'un biofilm. Certaines microcolonies ont une structure conique, d'autres ont l'aspect d'un champignon.Les flèches noires représentent le flux de fluide à travers les cavités d'eau et même par-dessous les microcolonies (Costerton, Lewandowski et al. 1995)

multicellulaire est indispensable à la survie de ces cellules. Un autre exemple est celui d'*Erwinia carotovora* qui synthétise des exoenzymes allant dégrader la paroi des cellules de la plante hôte. Cette production est sous le contrôle de deux types d'AHLs et cela permet d'assurer aux individus d'une population une attaque concertée efficace sans risque d'utilisation inutile d'exoenzymes.

De plus, cette collectivité protège la population contre une attaque que ne pourrait soutenir une cellule individuelle. Ainsi, les Actinomycètes produisent des antibiotiques dont la synthèse est régulée par des signaux de type γ-butyrolactones. Ces antibiotiques vont leur permettre d'éliminer des compétiteurs potentiels soit pour la niche écologique, soit pour la nourriture. Il existe d'autres exemples, comme la production de carbapenem par *Erwinia carotovora*, qui est régulée par une AHL (McGowan *et al.*, 1995), ou encore tels que la production de peptides antimicrobiens par les *Gram-positives* (Kleerebezem *et al.*, 1997).

Un autre moyen qu'ont trouvé les bactéries pour coloniser les milieux les plus variés et se protéger d'attaques extérieures (bactéricides,...), est leur organisation en biofilms. Des biofilms de communautés bactériennes mixtes et d'espèces individuelles telles que *Pseudomonas aeruginosa* se développent sur des surfaces solides exposées à un flux continuel de nutriments. Ils forment d'épaisses couches (environ 100 µm) constituées de structures ayant l'aspect de champignons et de piliers, séparées par des cavités remplies d'eau (figure 4). Les bactéries sont enfoncées dans ces structures composées d'une matrice de polysaccharide extracellulaire ou glycocalyx (Costerton *et al.*, 1995). Il a été montré qu'un signal intercellulaire intervenant dans le Quorum Sensing chez *Pseudomonas aeruginosa,* contrôle le développement des biofilms de cette espèce (Davies *et al.*, 1998).

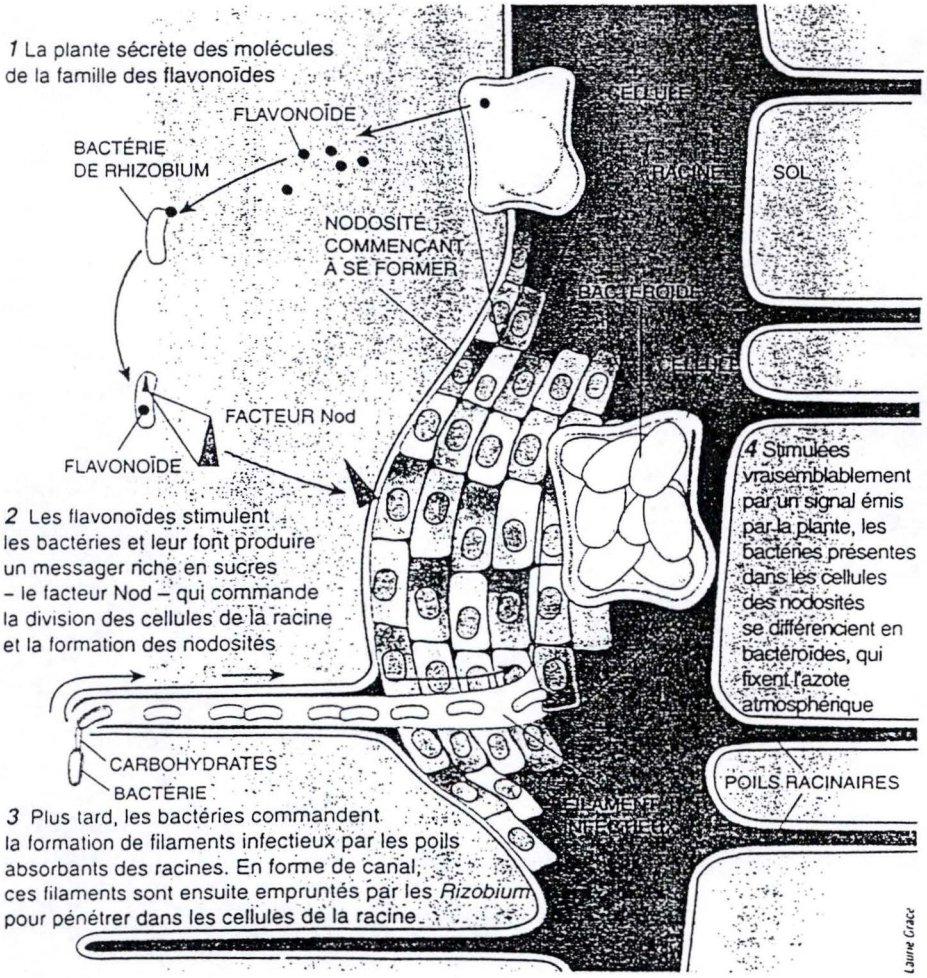
1.1.3.La sporulation

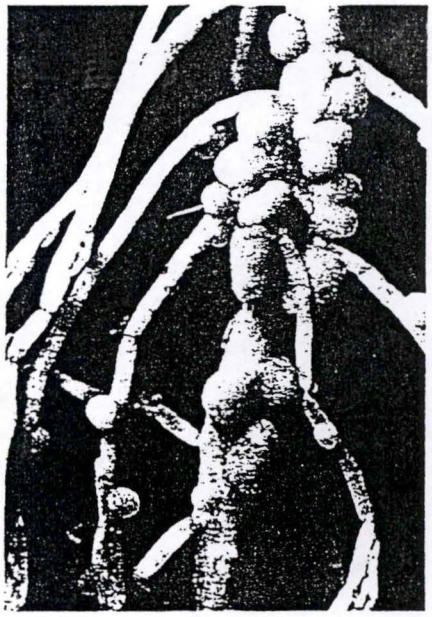
Elle implique une régulation multicellulaire et a été bien étudiée chez des organismes tels que *Bacillus subtilis* (Grossman, 1995) ou *Streptomyces coelicolor* (Chater and Losick, 1997). Chez cette dernière espèce, la sporulation est souvent associée à la morphogenèse multicellulaire (agrégation des cellules en une structure macroscopique).

1.1.4.L'échange de matériel génétique

Une population bactérienne de *Bacillus subtilis* peut développer des sous-populations compétentes pour l'échange d'ADN, permettant une nouvelle prolifération et une augmentation des chances de survie[1]. Le développement de cette compétence est un processus multicellulaire impliquant des signaux intercellulaires.

[1] Par des mécanismes d'adaptation tels que la recombinaison





LES NODOSITÉS qui se forment sur les racines du pois *(ci-dessus)* contiennent des bactéries du genre *Rhizobium*, qui fixent l'azote atmosphérique (N_2) et le convertissent en une forme utilisable par la plante. En échange, la plante abrite et nourrit les bactéries. La formation de ces nodosités et leur colonisation par les bactéries résultent de communications chimiques entre les racines et les bactéries *(schéma de gauche)*.

Figure 5 : Les échanges de molécules entre des bactéries du genre *Rhizobium* et une légumineuse (Losick and Kaiser 1997)

Les cellules compétentes de *Bacillus subtilis* peuvent intégrer l'ADN de n'importe quelle source (Solomon and Grossman, 1996).

1.1.5.Les bactéries fixatrices d'azote

Les interactions entre des bactéries et des organismes supérieurs peuvent aussi être considérées comme un comportement multicellulaire. Ainsi, certaines bactéries du sol partagent une relation symbiotique avec des légumineuses (par exemple les pois, les haricots, les lentilles,...). Elles contribuent aux échanges d'azote entre le sol et l'atmosphère. En effet, ces bactéries du genre *Rhizobium* sont parmi les rares organismes à pouvoir incorporer l'azote atmosphérique (N_2). Elles fournissent ensuite aux plantes auxquelles elles sont associées l'azote sous une forme assimilable (l'ammoniaque : NH_3). En échange, les plantes leur donnent la nourriture nécessaire sous la forme de sucres. Ces échanges se font dans des structures spécialisées de la racine appelées nodosités. Elles résultent de l'interaction plante-*Rhizobium*. Ce processus de fabrication implique une communication moléculaire à deux sens qui commence avant même que la bactérie ne soit rentrée en contact avec la plante (figure 5). Ainsi, une molécule de la famille des flavonoïdes produite par la plante, va aller stimuler certaines bactéries se trouvant dans le sol. Ces dernières répondent alors par une production de messagers riches en sucre appelés facteurs Nod qui vont commander la formation des nodosités. D'autres étapes successives où plante et bactéries se stimulent mutuellement aboutiront à l'installation finale des bactéries[1] dans les nodosités (Losick and Kaiser, 1997).

[1] Une fois installées, les bactéries grossissent, s'arrondissent et se transforment en centres de fixation de l'azote. Elles sont appelées alors **bactéroïdes**

2. Le Quorum Sensing des bactéries Gram-négatives

2.1 Généralités

2.1.1.Définition

Le Quorum Sensing est un phénomène contrôlé par des signaux extracellulaires qui permet à la bactérie de percevoir quand l'unité de population minimale ou « quorum » de bactéries est atteint et de répondre alors par une activation de la machinerie transcriptionnelle. Ces signaux extracellulaires sont les N-acyl-L-homosérines lactones (AHLs ou N-acyl-L-HSLs) chez les espèces *Gram-négatives* uniquement (Fuqua *et al.*, 1994). La structure de ces molécules est détaillée au point 2.2.1.1.

2.1.2.La découverte du système LuxR/LuxI

La première observation de ce phénomène fut effectuée en 1970 par des biologistes marins qui étudiaient des bactéries marines luminescentes telles que *Vibrio fischeri* à l'Université Harvard. Kenneth Nealson et John Woodland Hastings détectèrent une variation de la luminescence produite par les bactéries de leurs cultures. De plus, elles n'émettaient que lorsqu'une densité de population suffisante était atteinte.

Ils savaient déjà que la luminescence était le résultat de réactions chimiques catalysées par une enzyme, la luciférase. Ils supposèrent que la production de cette enzyme était commandée par un signal chimique que les bactéries s'échangeaient entre-elles et non pas par un mécanisme propre à chaque cellule. Leur hypothèse était la suivante : ce signal chimique, nommé autoinducteur[1] diffusait à l'intérieur de la cellule cible et quand sa concentration était suffisante, il allait activer l'expression des gènes codant les enzymes et d'autres protéines intervenant dans la luminescence. Bien que cette hypothèse ne fût pas accueillie favorablement, elle fut amplement démontrée par la suite (Losick and Kaiser, 1997).

Vibrio fischeri est devenu l'organisme de référence utilisé pour l'étude génétique et biochimique du Quorum Sensing. Il a été démontré que deux gènes jouent un rôle capital : *luxI* qui code la synthase de l'autoinducteur (LuxI) et *luxR* qui code le régulateur de la transcription des gènes de la luminescence (LuxR), LuxR étant activé par l'autoinducteur (Engebrecht and Silverman, 1984). L'autoinducteur a

[1] Le terme d'**autoinducteur** permit de donner un premier nom à ce phénomène : « **Autoinduction** » qui devint plus tard « **Quorum Sensing** »

N-(butanoyl)-L-homoserine lactone	homoserine lactone non-substituée en position 3+chaîne acyl de 4C	BHL
N-(hexanoyl)-L-homoserine lactone	homoserine lactone non-substituée en position 3+chaîne acyl de 6C	HHL
N-(octanoyl)-L-homoserine lactone	homoserine lactone non-substituée en position 3+chaîne acyl de 8C	OHL
N-(decanoyl)-L-homoserine lactone	homoserine lactone non-substituée en position 3+chaîne acyl de 10C	DHL
N-(dodecanoyl)-L-homoserine lactone	homoserine lactone non-substituée en position 3+chaîne acyl de 12C	dDHL
N-3(-oxo-butanoyl)-L-homoserine lactone	homoserine lactone substituée en position 3+chaîne acyl de 4C	OBHL
N-3(-oxo-hexanoyl)-L-homoserine lactone	homoserine lactone substituée en position 3+chaîne acyl de 6C	OHHL
N-3(-oxo-octanoyl)-L-homoserine lactone	homoserine lactone substituée en position 3+chaîne acyl de 8C	OOHL
N-3(-oxo-decanoyl)-L-homoserine lactone	homoserine lactone substituée en position 3+chaîne acyl de 10C	ODHL
N-3(-oxo-dodecanoyl)-L-homoserine lactone	homoserine lactone substituée en position 3+chaîne acyl de 12C	OdDHL

Tableau 2 : Les abréviations correspondant à différentes phéromones

été identifié comme étant la OHHL (Eberhard *et al.*, 1981) (tableau 2). A faible densité cellulaire, *luxI* est transcrite à un taux basal et la phéromone (VAI : *Vibrio* AutoInducer) diffuse passivement en dehors de la cellule selon un gradient de concentration. A haute densité cellulaire, elle s'accumule (les concentrations internes et externes s'égalisant alors). Une concentration de l'ordre de 10 nM est suffisante pour qu'elle puisse interagir avec LuxR qui va alors activer la transcription des gènes de la luminescence ainsi que celle de *luxR* et *luxI* eux-mêmes, créant un rétrocontrôle positif (Fuqua *et al.*, 1994). LuxR est localisé au niveau de la membrane cellulaire, du côté cytoplasmique. Il peut ainsi répondre à la concentration environnementale en autoinducteur.

Nous pouvons nous demander quel avantage tire *Vibrio fischeri* de ce système. La réponse est en relation avec certains poissons et calamars. En effet, *Vibrio fischeri* est le symbionte spécifique de ces animaux. Le calamar *Euprymna scolopes*, par exemple, laisse proliférer ces bactéries dans son organe lumineux où la densité cellulaire est élevée (10^{10} à 10^{11} cellules/ml), elles y sont donc luminescentes. Tandis que dans l'eau de mer, elles se retrouvent à une densité de moins de 10^2 cellules/ml et ne sont pas luminescentes (Fuqua *et al.*, 1994). Le calamar, qui est un chasseur nocturne, profite de cette lumière car elle lui permet de rester « invisible » aussi bien pour ses proies que pour ses prédateurs nageant sous lui; en effet, cette lumière efface l'ombre que la lune produit. *Vibrio fischeri* trouve aussi un bénéfice à ces interactions puisque l'organe lumineux lui fournit protection et nourriture. Le système d'autoinduction permet d'éviter un gaspillage d'énergie pour la production de lumière car elle n'est pas utile quand la bactérie se situe en dehors de l'organe lumineux (Losick and Kaiser, 1997).

2.1.3.Les systèmes de régulation homologues

Au début des années 1990, plusieurs groupes scientifiques découvrirent des systèmes de régulation de type LuxR/LuxI chez d'autres bactéries. La diversité des espèces utilisant l'autoinduction et la similarité biochimique et génétique de ces systèmes, indiquèrent que ce phénomène était un mécanisme de régulation commun aux bactéries *Gram-négatives*. Des homologues de LuxR furent découverts d'abord chez *Pseudomonas aeruginosa* (Gambello and Iglewski, 1991) et *Agrobacterium tumefaciens* (Zhang *et al.*, 1993) et il fut montré que plusieurs espèces produisaient des OHHL tout comme *Vibrio fischeri* (Greenberg, 1997). Le fonctionnement de ces mécanismes de régulation n'est pas toujours aussi simple que celui de *Vibrio fischeri*. En effet, certaines bactéries ont plusieurs systèmes de Quorum Sensing comme, par exemple, *Pseudomonas aeruginosa* qui en a au moins deux. Un de ces systèmes utilise la OdDHL (AHL substituée en position 3) et l'autre la BHL (AHL non-substituée en position 3) (Fuqua *et al.*, 1994). Il existe aussi des bactéries telles que *Vibrio harveyi*, où Quorum Sensing et système de régulation à deux composantes sont liés (Swift *et al.*, 1994). La famille LuxR regroupe les homologues de LuxR. Leur identité de séquence protéique avec LuxR varie entre 18 et 27% (figure 6). La

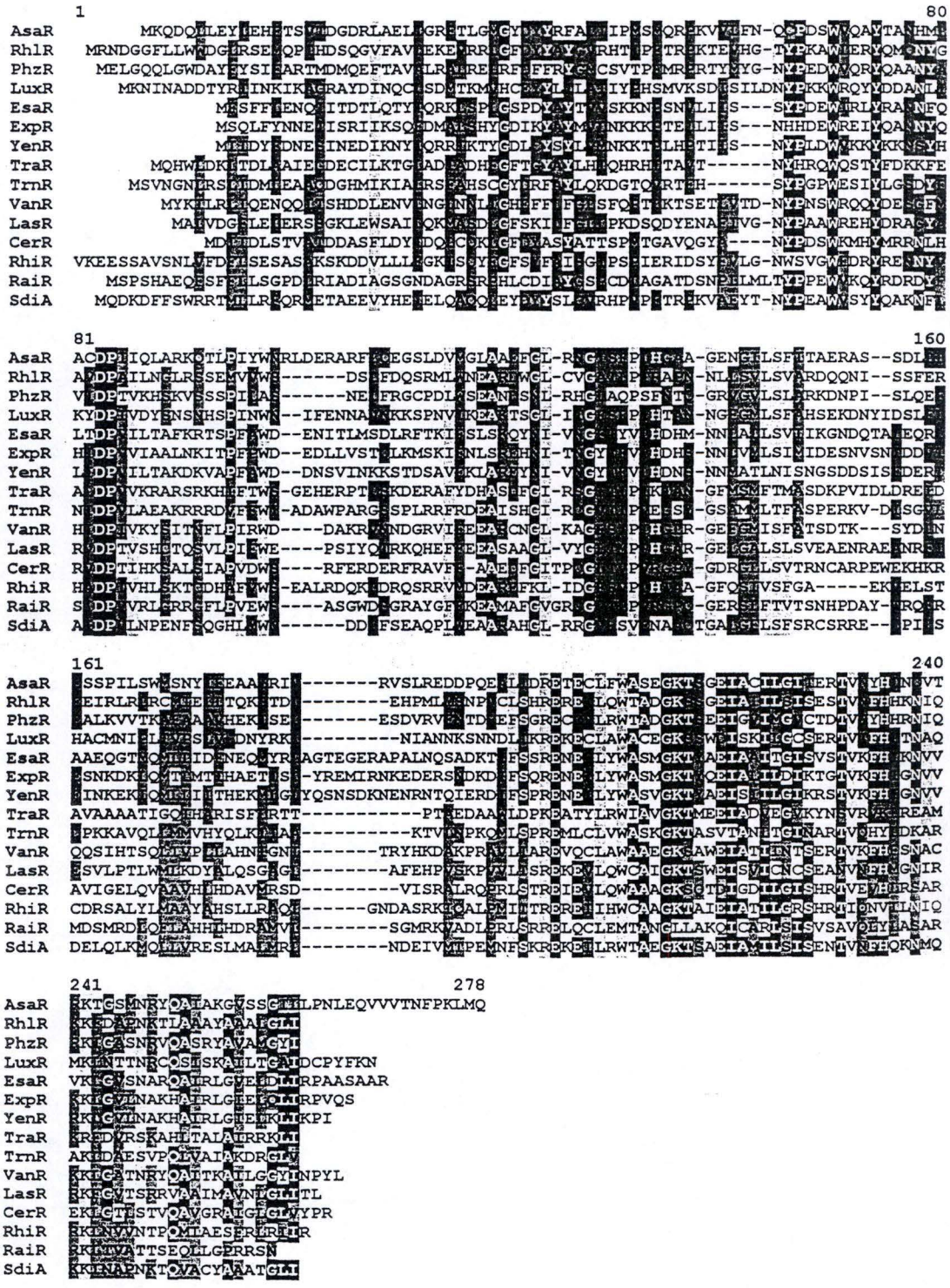


Figure 6 : Alignement d'homologues de LuxR. Cet alignement a été réalisé par l'algorithme d'alignement multiple Clustal. Les surlignages en noir représentent des résidus invariants et les plus clairs représentent des résidus dont la similarité entre eux est plus faible (Stevens and Greenberg 1999).

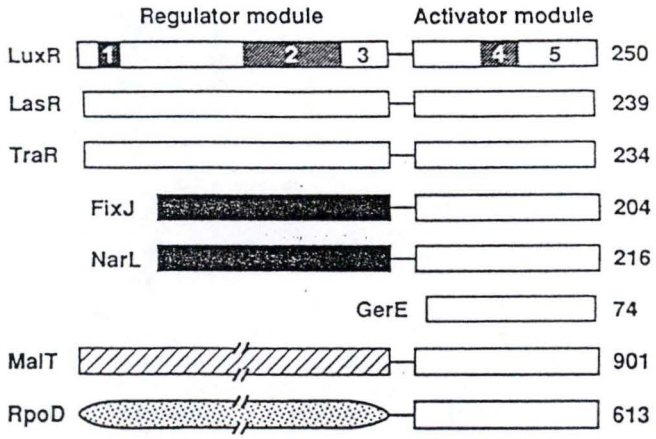


Figure 7 : Diagramme schématique de LuxR (les zones 1 à 5 correspondent à des régions aux fonctions connues se retrouvant chez les autres membres de la famille LuxR -> cfr. point 2.2.2.2.) et des membres représentatifs de la superfamille LuxR. Le premier domaine de chaque protéine est appelé le domaine régulateur et le second, le domaine activateur (Fuqua, Winans et al. 1994).

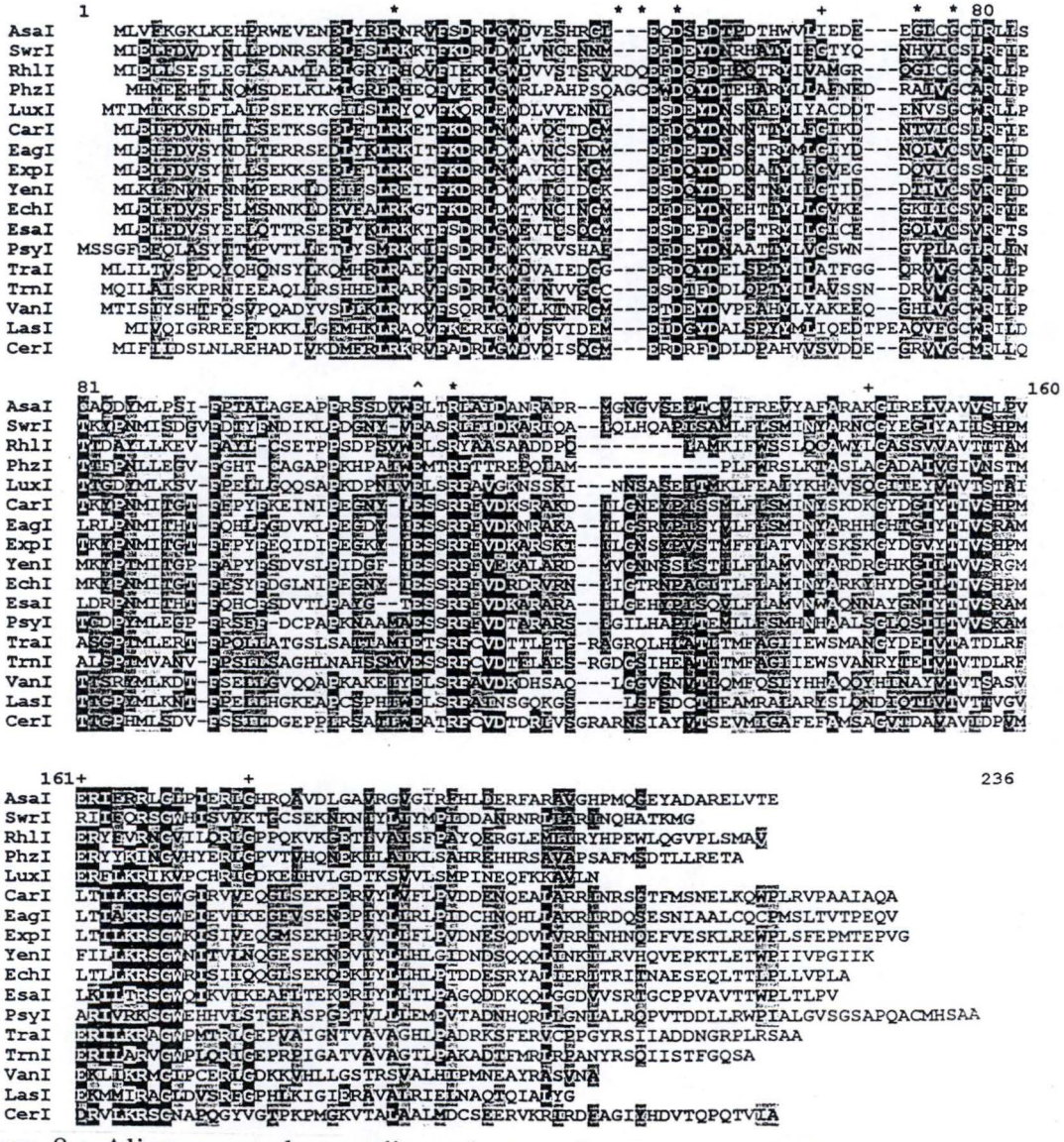


Figure 8 : Alignement de protéines de type LuxI donné par l'algorithme d'alignement multiple Clustal. Les surlignages en noir représentent les résidus invariants et les plus clairs représentent les résidus dont la similarité entre eux est plus faible. (*) La mutation de ces résidus résulte en une perte de fonction de LuxI et RhlI, (+) de LuxI seule, (^) de RhlI seule (Fuqua and Eberhard 1999).

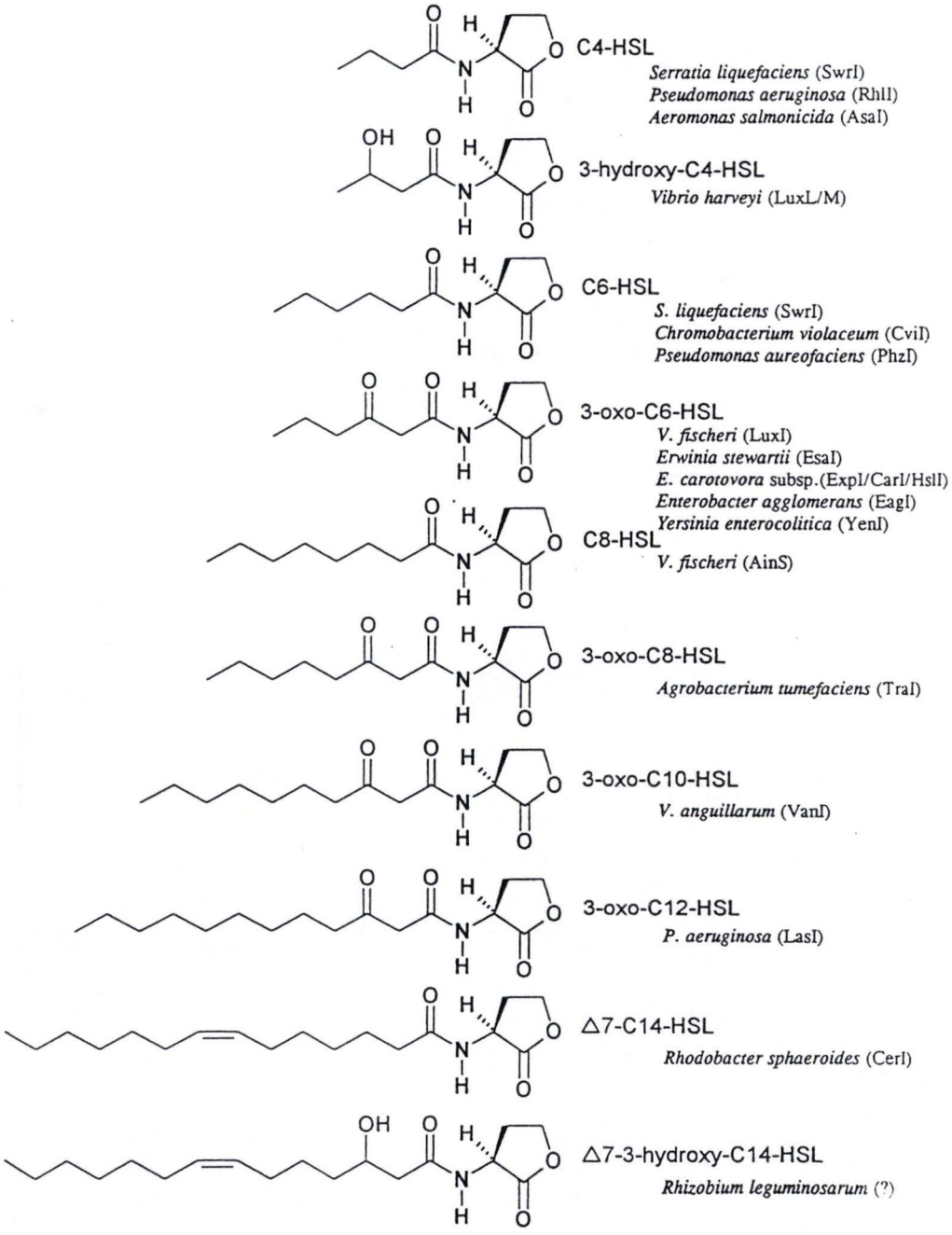


Figure 9 : Comparaison de la structure de différentes AHLs. Le nom des synthases correspondantes est placé entre parenthèses derrière le nom de l'organisme produisant la phéromone (Fuqua and Eberhard 1999).

similarité, pour chaque homologue, est répartie dans des régions s'alignant avec deux portions de LuxR : une portion du domaine N-terminal supposée interagir avec l'AHL et une portion du domaine C-terminal supposée interagir avec l'ADN (le domaine HTH) (Dunlap, 1997). Cette famille est rattachée à la superfamille LuxR dont les membres exhibent de l'homologie avec LuxR uniquement avec son domaine C-terminal. Cette superfamille comprend aussi des familles telles que MalT, FixJ, NarL et GerE (figure 7). Des homologues de LuxI ont également été identifiés chez d'autres bactéries. Elles partagent entre 25 et 35% d'identité de séquence protéique avec LuxI (figure 8) et elles dirigent la synthèse d'AHLs ayant une chaîne acyl saturée ou insaturée de 4 à 14 carbones. Chacune de ces AHLs a un groupe hydroxyle, un groupe carbonyle ou un hydrogène en position 3 (le troisième carbone compté à partir du lien amide). Le caractère commun de ces molécules est le noyau homosérine lactone alors que la chaîne acyl de longueur et de biochimie variables fournit la spécificité (figure 9).

2.1.4.Quelques exemples

2.1.4.1.*Agrobacterium tumefaciens*

Agrobacterium tumefaciens, un pathogène de plantes, provoque la galle du collet chez la plante hôte en transférant des fragments d'ADN oncogènes de plasmides Ti (« Tumor Inducing ») vers le noyau de la cellule végétale cible. Certains de ces gènes codent pour des opines qui, une fois sécrétées, serviront de nourriture à *Agrobacterium tumefaciens*.

Les plasmides Ti dirigent leur propre transfert conjugatif entre les bactéries, au moyen des gènes *tra* (figure 10). Ce transfert conjugatif ne se fait que dans les tumeurs, symptômes de la maladie, ou en présence d'opines exogènes. La conjugaison requiert une densité cellulaire élevée et est contrôlée par une paire de protéines de type LuxR/LuxI, appelées TraR et TraI. TraR active l'expression des gènes *tra* de plasmides Ti et est similaire à LuxR, non seulement pour le domaine C-terminal mais pour l'entièreté de la séquence protéique. TraI, l'homologue de LuxI, synthétise une phéromone diffusible de type OOHL. Le complexe TraR-AAI (*Agrobacterium* AutoInducer) active l'expression d'au moins deux opérons *tra*, responsables du transfert conjugatif de plasmides Ti entre bactéries, ainsi que celle de *TraI* et *TraR* eux-mêmes. Ceci crée un rétrocontrôle positif semblable à celui de *Vibrio fischeri*.

Deux types d'opines sont produits : l'octopine dont l'action touche les plasmides Ti de type octopine et l'agrocinopine dont l'action touche les plasmides Ti de type nopaline (Fuqua *et al.*, 1994). Dans un plasmide Ti de type octopine, la protéine OccR transcrite à partir de ce plasmide va, en présence d'octopine, activer la

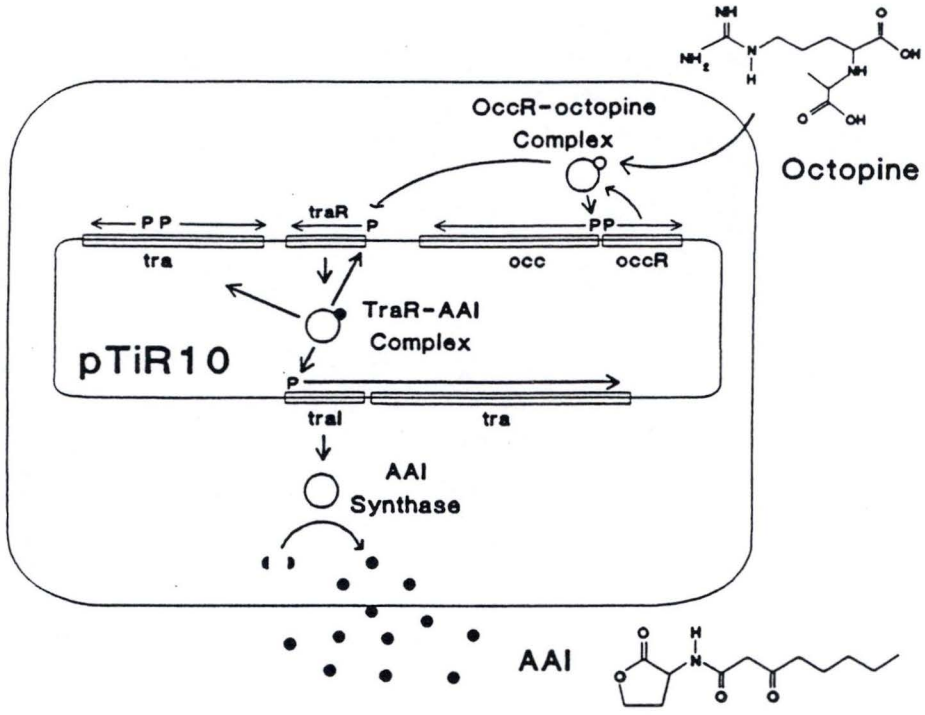


Figure 10 : Modèle régulatoire du contrôle de l'expression des gènes *tra* chez *A. tumefaciens* . C'est la régulation des plasmides Ti de type octopine qui est montrée ici. Dans les plasmides Ti de type nopaline, l'agrocinopine inactive un répresseur, AccR (Dunlap 1997).

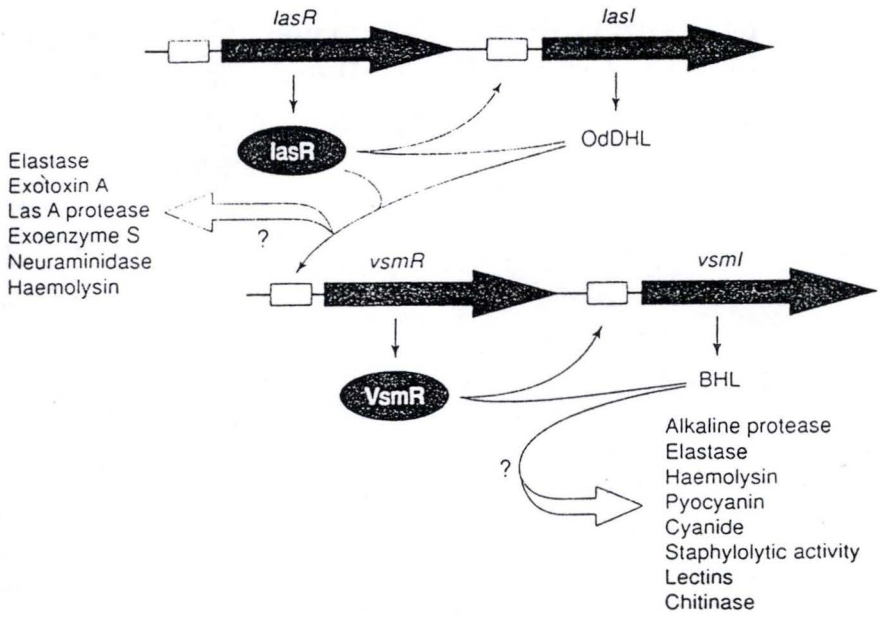


Figure 11 : Arrangement hiérarchique des circuits du Quorum Sensing chez *P. aeruginosa* . Les interactions potentielles sont marquées par '?' (Swift, Throup et al. 1996).

transcription des locus *Occ*[1] et *TraR*. TraR activera alors l'expression des gènes *tra* en présence d'AAI dont la synthèse dépend de TraI. Une voie de régulation semblable se retrouve pour les plasmides Ti de type nopaline. Seulement, la protéine AccR transcrite à partir de ce plasmide, va réprimer la transcription des gènes *tra* en absence d'agrocinopine. Donc, ce processus exige la production d'opines par la plante ET une densité cellulaire suffisante puisque le Quorum Sensing contrôle le transfert (Fuqua *et al.*, 1994).

2.1.4.2.*Pseudomonas aeruginosa*

Le pathogène opportuniste *Pseudomonas aeruginosa* infecte les tissus blessés et cause des maladies chroniques des voies respiratoires chez les immunodéprimés ou les malades atteints de mucoviscidose. Toute une série de facteurs de virulence interviennent lors de l'infection, incluant par exemple, les protéases spécifiques de l'élastine, LasA et LasB. Un régulateur de la famille LuxR a été isolé et désigné sous le nom de LasR. Il régule l'expression de LasA mais aussi celle de aprA (protéase alkaline A) et de toxA (exotoxine A). Une synthase homologue à LuxI (LasI) a également été isolée. Elle dirige la synthèse de OdDHL appelée aussi PAI (*Pseudomonas* AutoInducer). A faible densité cellulaire, cette phéromone n'est pas à une concentration suffisante pour activer LasR. Par contre, à densité cellulaire élevée, elle en sera capable et cela aboutira à l'expression des gènes de virulence. La dégradation protéolytique sera alors effective (Fuqua *et al.*, 1994).

Un autre système de Quorum Sensing a été découvert chez cette espèce. Il fait intervenir le régulateur transcriptionnel VsmR et la synthase VsmI qui produit une AHL non-substituée en position 3 (BHL). Le mécanisme est le même que celui de l'autre système et l'activation de VsmR par la BHL aboutit également à l'expression de plusieurs facteurs de virulence (protéases alkalines, élastases, lectines, chitinases,...). Les deux systèmes sont interdépendants et semblent former une cascade hiérarchique de Quorum Sensing (figure 11). Ainsi, le complexe LasR-OdDHL est requis pour l'expression de *VsmR* et le complexe VsmR-BHL réprime l'expression de *LasR* (Fuqua *et al.*, 1994).

2.1.4.3.*Erwinia carotovora*

Erwinia carotovora est un pathogène de végétaux et colonise les tissus vasculaires de la plante hôte. Elle produit des enzymes extracellulaires qui vont aller dégrader les parois cellulaires de la plante (polygalacturonases, cellulases, protéases,...). Cette production d'exoenzymes est responsable de la macération des tissus et elle est nécessaire à la bactérie pour qu'elle puisse se propager dans la plante.

---

[1] Les gènes *Occ* sont responsables du catabolisme des **octopines**. Leurs équivalents dans les **plasmides Ti** de type nopaline, pour le catabolisme des **agrocinopines**, sont les gènes *Acc*.

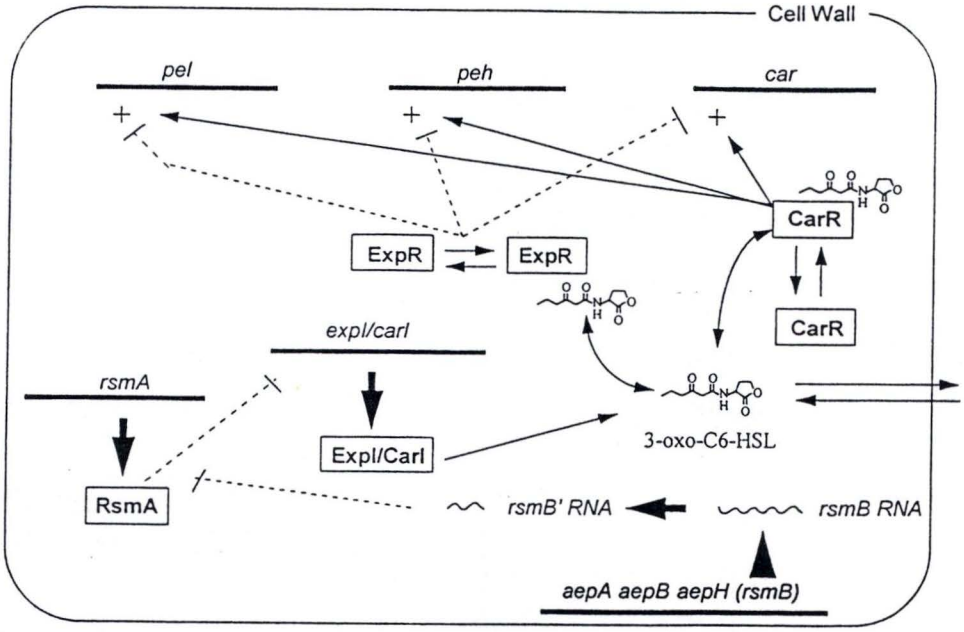


Figure 12 : Modèle de la régulation des facteurs de virulence chez *E. carotovora* . Les facteurs de virulence incluent la pectate lyase (Pel), la polygalacturonase (Peh) et le carbapenem qui est un antibiotique (Car) (Pierson III, Wood et al. 1999).

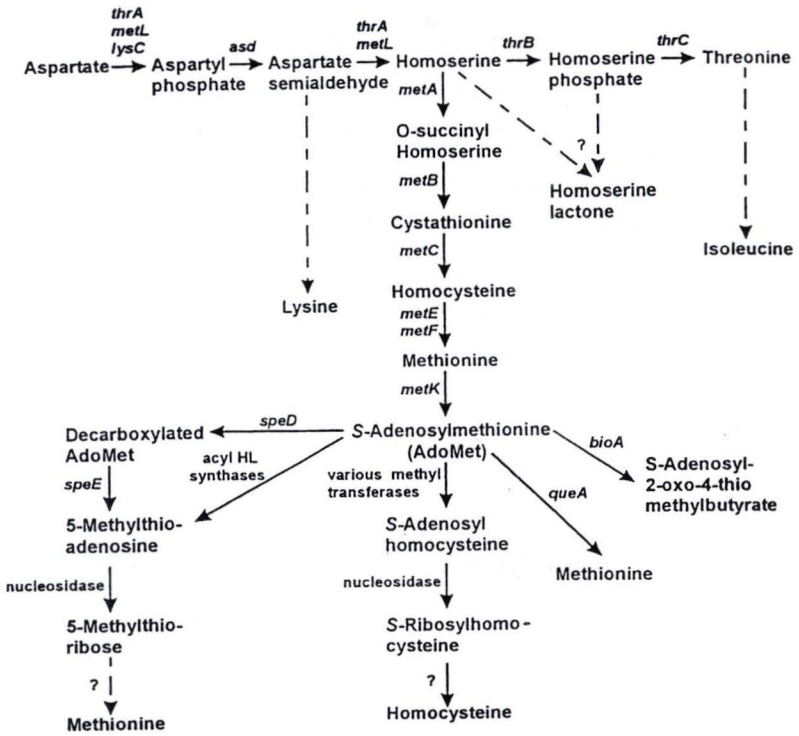


Figure 13 : Voie de synthèse biochimique méthionine-lysine-thréonine (Fuqua and Eberhard 1999).

Il existe plusieurs systèmes de régulation qui ont été identifiés dans plusieurs souches d'*Erwinia carotovora* (figure 12). Un de ces systèmes fait intervenir ExpR et ExpI. ExpI est responsable de la synthèse d'une phéromone, la OHHL. Le rôle du régulateur ExpR n'est pas encore très bien défini, mais il pourrait fonctionner en tant que répresseur de l'expression d'exoenzymes en titrant la OHHL (McGowan *et al.*, 1995).

Un autre système, le système CarR-CarI, contrôle la biosynthèse d'un antibiotique, le carbapenem. Cette production est synchronisée avec celle d'exoenzymes, ce qui laisse supposer qu'elle permet de réduire la population des compétiteurs éventuels. CarI synthétise la OHHL tout comme ExpI. Elle sert de co-inducteur à CarR pour l'activation de la transcription des gènes de la production du carbapenem et celle des gènes de la biosynthèse d'exoenzymes (McGowan *et al.*, 1995).

Deux autres systèmes de régulation, RsmA et RsmB (Aep A, B, H), sont reliés aux deux circuits précédents. RsmA réduit la stabilité des mRNA de gènes tels que *ExpI, CarI, pel, peh* en se liant à ces messagers ; leur nombre est ainsi réduit. Les protéines AepA, B, H (RsmB) sont nécessaires à la production des exoenzymes. AepA et B activent la transcription de gènes encodant des enzymes extracellulaires en réponse à des composés produits par la plante. AepH augmente l'activité transcriptionnelle d'AepA et B. RsmA est régulée par RsmB et cette régulation requiert le facteur sigma RpoS. RsmB exerce un effet négatif sur l'expression de *RsmA* ou sur la fonction de RsmA. Ceci permet alors de lever l'inhibition de l'expression de gènes tels que celui de l'AHL-synthase et d'avoir un taux suffisant d'AHLs pour une production de carbapenem et d'exoenzymes (cfr figure 12) (Pierson III *et al.*, 1999).

2.1.4.4.*Escherichia coli*

Un des premiers homologues de LuxR à avoir été découvert est une protéine d'*E.coli* impliquée dans la division cellulaire et dénommée SdiA. Elle active la transcription de l'opéron *ftsQAZ* dont les produits sont responsables de la septation cellulaire. Étonnement, aucun homologue de LuxI n'a encore été découvert chez *E.coli*. Cette régulation par un homologue de LuxR suggère que la division cellulaire de cette espèce est un phénomène dépendant de la densité (Fuqua *et al.*, 1994).

2.2. Aspects moléculaires

2.2.1.LuxI et ses homologues

2.2.1.1.Variations dans la structure des N-acyl-L-HSLs

Les N-acyl-L-HSLs des bactéries *Gram-négatives* sont synthétisées par les N-acyl-L-HSL synthases qui partagent de l'homologie avec LuxI. Certaines bactéries utilisent la même phéromone que *Vibrio fischeri*, c'est-à-dire la OHHL, tandis que d'autres emploient une phéromone avec une chaîne acyl de longueur différente, à un état d'oxydation différent et à un degré de saturation différent (cfr. figure 9). Toutes ces N-acyl-L-HSLs sont composées d'une chaîne acyl d'une longueur allant de 4 à 14 carbones, liée à un noyau homosérine lactone par un lien amide. Un premier site de variation est le carbone en position 3 sur la chaîne acyl : il peut y avoir un groupe carbonyl, un groupe hydroxyl où il peut être complètement réduit. Il existe aussi des exemples pour lesquels la chaîne acyl est insaturée à un endroit (Puskas *et al.*, 1997) (cfr. figure 9). Les différents états d'oxydation en position 3, ainsi que le lien insaturé éventuel, imposent une certaine structure à la phéromone.

2.2.1.2.Les substrats de la N-acyl-L-HSL synthase

Les substrats de la N-acyl-L-HSL synthase sont la S-adénosylméthionine (AdoMet ou SAM) et l'acide gras approprié conjugué à une protéine porteuse d'acyl (ACP=Acyl Carrier Protein).

L'homosérine est présente dans de nombreuses bactéries en tant qu'intermédiaire de la voie de synthèse biochimique méthionine-lysine-thréonine (figure 13).Nous pourrions penser que l'homosérine des N-acyl-L-HSLs est prélevée du pool cellulaire d'homosérines ou d'homosérines lactones mais des expériences sur des cultures cellulaires de *Vibrio fischeri* ont montré que le substrat le plus effectif est la méthionine ou la S-adénosylméthionine (Eberhard *et al.*, 1991). L'AdoMet est la première source de groupements méthyl pour plusieurs réactions cellulaires. Elle est synthétisée à partir de méthionine et d'ATP par la protéine MetK (AdoMet synthase) (cfr. figure 13). Les N-acyl-L-HSL synthases n'ont pas un motif de liaison à l'AdoMet discernable et ne partagent pas d'homologie avec les autres enzymes utilisant l'AdoMet comme substrat. Les N-acyl-L-HSL synthases ont peut-être développé un nouveau mécanisme d'interactions avec l'AdoMet ou alors le site actif est très peu similaire au site actif de liaison habituel. Cette dernière hypothèse est la plus probable car les AHL synthases n'utilisent pas seulement le méthyl mais une bonne partie de l'AdoMet.

La structure de la chaîne acyl des N-acyl-L-HSLs suggère que la synthèse de ces composés recrute des acides gras du métabolisme des lipides (Eberhard *et al.*,

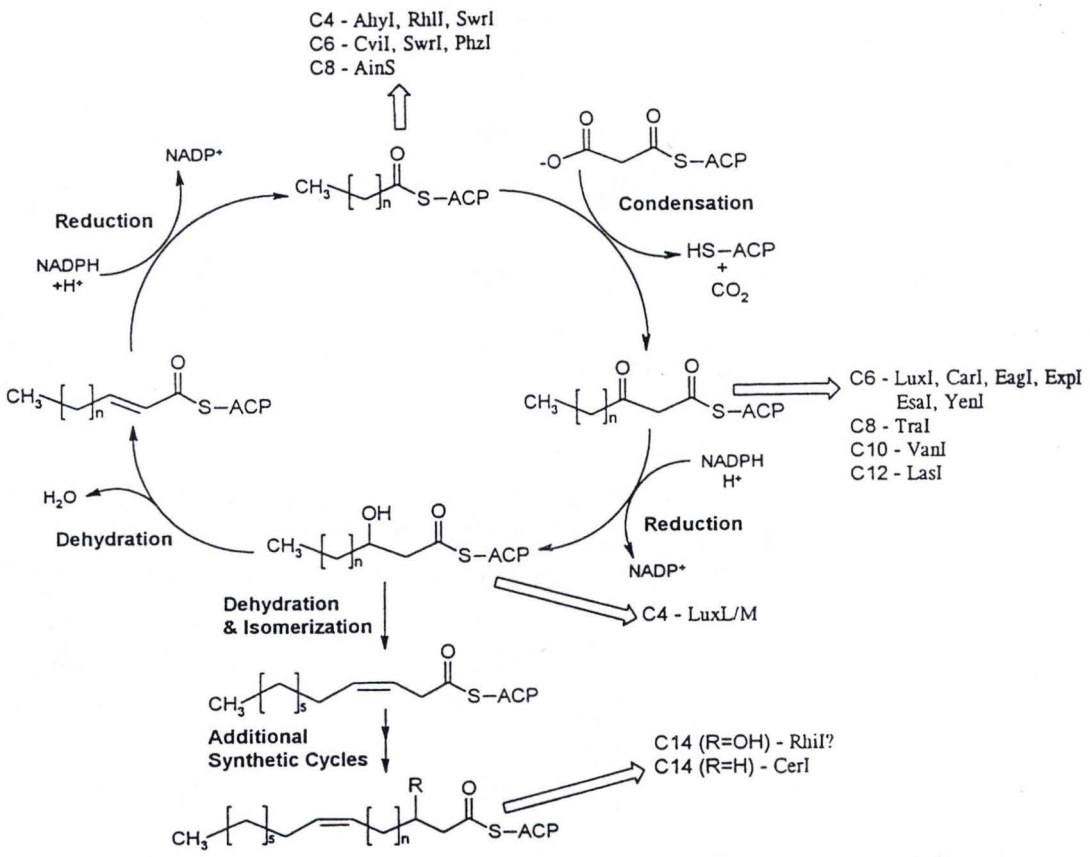


Figure 14 : Cycle de synthèse des différents types d'acyl-ACP. Les flèches larges indiquent les points où les AHL synthases peuvent recruter leur substrat (acyl-ACP) (Fuqua and Eberhard 1999).

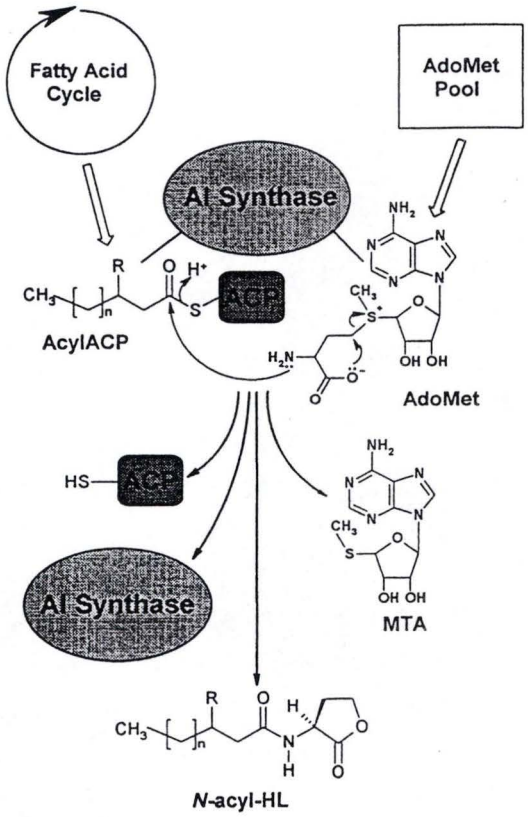


Figure 15 : Mécanisme catalytique de synthèse des AHLs par les protéines de type LuxI. MTA, 5'-méthylthioadénosine (Fuqua and Eberhard 1999).

1981). La biosynthèse des acides gras se réalise par addition pas à pas d'unités de 2 carbones transférés d'une unité malonyl-ACP vers la chaîne acyl naissante. Celle-ci est liée de façon covalente à une ACP *via* un lien thioester (figure 14). Lors de la biodégradation, les acides gras sont activés par la conjugaison avec une coenzyme A et sont soumis ensuite aux différentes étapes de la β-oxydation. Il existe des preuves de la transestérification du conjugué acyl-CoA directement sur l'ACP, fournissant ainsi un lien entre la dégradation et la biosynthèse d'acides gras. La biosynthèse d'acides gras est requise dans la réaction pour pouvoir toujours fournir le substrat approprié à une ACP.

2.2.1.3.La synthèse de la N-acyl-L-HSL

Le groupement amine de l'AdoMet permet une attaque nucléophile du groupement carbonyl (figure 15). Il n'est pas encore certain que cette attaque nucléophile se fasse sur le conjugué acyl-ACP lui-même. Il se peut aussi qu'elle se fasse sur un composé intermédiaire entre l'acyl-ACP et la N-acyl-L-HSL. La réaction se poursuit en liant la chaîne acyl à l'AdoMet *via* un lien amide. La lactonisation (cyclisation) de l'intermédiaire acyl-AdoMet génère la N-acyl-L-HSL et la 5'-méthylthioadénosine (cfr. figure 15). L'ordre de catalyse de toutes ces étapes est encore en cours d'étude et nous ne savons pas si la lactonisation est dirigée enzymatiquement ou si elle se produit spontanément par la formation d'un lien amide. Il est intéressant de noter que les N-acyl-L-HSL synthases ont de 10 à 100 fois plus d'affinité pour les acyl-ACP que pour les AdoMet. Cette différence de K_m (coefficient d'affinité) s'explique par le fait que, dans des cellules en croissance, le taux d'AdoMet est élevé et donc l'affinité pour ce substrat peut être basse. Par contre, le pool d'acyl-ACP requis représente uniquement une petite fraction du pool d'acyl-ACP total et donc un K_m peu élevé pour ce substrat permet la synthèse d'une quantité adéquate de N-acyl-L-HSL sans réduire exagérément les pools spécifiques d'acides gras dans la cellule.

2.2.1.4.Les interactions enzyme-substrat

Le site actif d'une N-acyl-L-HSL synthase doit être adapté à deux substrats différents : l'AdoMet et la chaîne acyl portée par une ACP , de longueur et de structure spécifique.

L'ACP est composée de quatre hélices α chargées négativement formant une structure en forme de bâtonnet. Les chaînes acyl de six carbones et moins restent cachées dans la fente située entre les hélices deux et trois. La reconnaissance du conjugué acyl-ACP spécifique par la phéromone synthase implique aussi bien les interactions avec la chaîne acyl (cachée ou non) que la perception de la variation de la structure de la protéine ACP causée par la portion cachée de la chaîne acyl. Finalement, il peut y avoir des interactions avec les enzymes de biosynthèse des acides gras qui dirigent l'étape précédant le recrutement pour la synthèse de N-acyl-L-

HSL. La phéromone synthase est donc spécifique à la longueur de la chaîne acyl du conjugué acyl-ACP et à la biochimie de cette chaîne. Les résidus requis pour la liaison à l'AdoMet devraient être conservés d'une N-acyl-L-HSL synthase à l'autre alors que ceux requis pour la liaison de l'acyl-ACP seraient spécifiques aux différents types d'acyl-ACP.

2.2.1.5.Les réactions annexes

Le produit primaire de ces N-acyl-L-HSL synthases active très bien les régulateurs transcriptionnels de type LuxR. Cependant, dans de nombreux cas, des composés additionnels sont synthétisés par les mêmes synthases à des concentrations détectables (Jones *et al.*, 1993). Ces produits annexes sont moins effectifs que les produits primaires. Ils pourraient jouer le rôle d'inhibiteurs compétitifs des protéines de type LuxR. Il existe au moins un exemple de ce type de compétition : chez *Vibrio fischeri,* LuxI synthétise la OHHL et une faible quantité de HHL alors qu'une autre N-acyl-L-HSL synthase appelée AinS[1] fabrique du OHL. Les souches portant une mutation nulle dans le gène *AinS* expriment les gènes de la luminescence à une densité cellulaire plus faible que ne le font les souches sauvages. Cet effet est aboli par addition de OHL synthétique. Nous pouvons en déduire que cette phéromone module l'activation de la luminescence par compétition avec la OHHL pour le régulateur LuxR.

2.2.1.6.Etudes structure/fonction de la phéromone synthase

Nous possédons peu d'informations sur la structure de ces protéines. Elles ont un rapport résidus hydrophobes/résidus hydrophiles typique et sont monomériques (Moré *et al.*, 1996). Des analyses de mutants ont déterminé les résidus essentiels à l'activité protéique. Deux régions ont ainsi été pointées (Hanzelka *et al.*, 1997). Une de celles-ci s'étend du résidu 24 au résidu 104 et l'autre, du résidu 133 au résidu 164 (cfr. figure 8). La plupart de ces mutations aboutissent à la perte complète de la fonction. Elles correspondent aux résidus conservés dans la famille LuxI. Les résidus les mieux conservés se trouvent dans la première région. C'est donc cette région-là qui pourrait interagir avec l'AdoMet. L'autre région, plus variable, pourrait intervenir dans la sélection du conjugué acyl-ACP. Les résidus conservés qui ont été désignés comme essentiels par des expériences de mutagenèse sont pour la plupart des résidus chargés. Cela suggère que la synthèse de N-acyl-L-HSL est médiée par des réactions acide-base dans le site actif. Par contre, les résidus conservés non-polaires ne sont pas essentiels aux réactions catalytiques. Ces derniers peuvent être impliqués dans la structuration du site actif (Fuqua and Eberhard, 1999).

[1] Cet enzyme n'est pas un homologue de LuxI

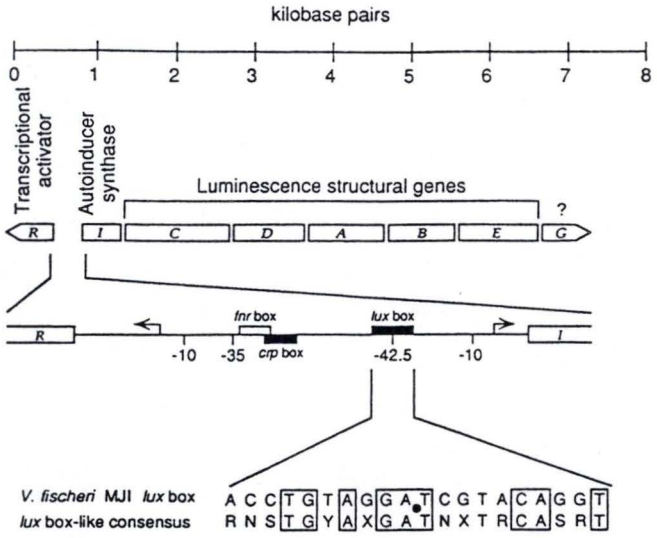


Figure 16 : Le cluster des gènes *lux* de *V. fischeri* , la région régulatoire et la *lux* box. En dessous, la séquence nucléique de la *lux* box et celle du consensus des équivalents de la *lux* box régulés par des homologues de LuxR dans d'autres organismes. N=A,T,C ou G ; S=C ou G ; X=N ou un trou dans la séquence (Stevens and Greenberg 1999).

2.2.1.7.Le contrôle de l'expression des gènes de type *luxI*

Ces gènes sont transcrits à un taux basal, mais les protéines de type LuxR activent cette transcription. Cela aboutit à un rétrocontrôle positif et à une augmentation de la production de phéromones. Ce rétrocontrôle positif amplifie encore le phénomène d'autoinduction et permet de répondre plus rapidement. Ce qu'il ne faut pas oublier, c'est que ce FeedBack positif n'est pas lui-même le processus d'autoinduction. Ce phénomène ne nécessite pas le contrôle positif du gène de la N-acyl-L-HSL synthase par un régulateur de type LuxR ; il requiert uniquement une expression à un taux basal de cette synthase puisque c'est une densité cellulaire élevée qui va augmenter la concentration en phéromone (Fuqua and Eberhard, 1999).

2.2.2.LuxR

2.2.2.1.L'opéron *lux* et la lux *box*

La caractérisation moléculaire du système de la luminescence a été rendue possible en 1983 grâce au clonage dans *E.coli* d'un fragment d'ADN de 9 kb qui regroupe tous les gènes impliqués dans le phénomène (Engebrecht *et al.*, 1983).

Tous ces gènes sont arrangés en deux unités transcriptionnelles séparées par 155 bp (figure 16). Une unité contient *luxR* et l'autre, l'opéron *luxICDABEG*. Les gènes *luxCDABEG* sont les gènes de la luminescence : *luxA* et *B* encodent les sous-unités α et β de la luciférase ; *luxC, D* et *E* encodent les composants du complexe « acides gras réductase » requis pour la synthèse des substrats de la luciférase ; *luxG* n'est pas requis pour la luminescence mais il pourrait encoder une flavine mononucléotide réductase fabriquant un substrat pour la luciférase. Dans la région de régulation *luxR-luxI* se trouve aussi le site de liaison de la protéine réceptrice de l'AMP cyclique (CRP) centré à environ –60 bp du site de départ de la transcription de *luxR*. La CRP est requise pour l'expression du gène *luxR*. Celui-ci est autorégulé. L'autorégulation peut être positive ou négative selon la concentration cellulaire de LuxR[1] et selon la présence ou l'absence d'éléments inhibiteurs dans l'ORF (Open Reading Frame) *luxD* (Shadel and Baldwin, 1991). D'autres facteurs sont connus pour intervenir dans la régulation des gènes *lux* : LexA, GroESL (aidant LuxR à se replier en une forme active), FNR (Adar *et al.*, 1992), (Ulitzer and Dunlap, 1995). Il a aussi été montré qu'une limitation du fer et de l'oxygène aboutit à une induction précoce de la luminescence (Ulitzer and Dunlap, 1995).

La partie clé de cette région régulatrice est la *lux* box qui fait 20 bp de long et est centrée à – 42 bp du site de départ de la transcription de *luxI*. Elle est caractérisée par une symétrie palindromique. La *lux* box est le site de liaison de LuxR. Une fois lié à l'ADN, LuxR peut alors activer la transcription. Des études de mutagenèse sur

[1] A haute concentration cellulaire, LuxR inhibe sa propre production tandis qu'à faible concentration cellulaire, LuxR l'active.

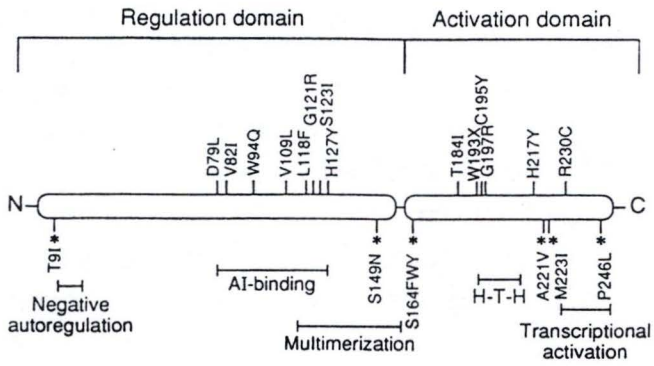


Figure 17 : Modèle des régions clés de LuxR. Les acides aminés indiqués par des chiffres et des lettres (les lettres représentant les substitutions) sont essentiels à l'activité autoinducteur-dépendante de LuxR. Si les résidus notés par une * sont substitués , il y a une activation autoinducteur-indépendante de LuxR (Stevens and Greenberg 1999).

ce site ont permis de montrer que la boîte minimale requise pour la liaison de LuxR a une longueur de 15 bp (Devine *et al.*, 1989). Une *lux* box consensus a pu être déterminée grâce à un alignement des box régulées par des homologues de LuxR (cfr. figure 16) (Fuqua *et al.*, 1994).

2.2.2.2.Le facteur transcriptionnel LuxR

LuxR est une protéine de 250 acides aminés. Des mutations ponctuelles ont été générées sur *luxR* pour déterminer quelles sont les régions essentielles à l'activité de LuxR (Shadel *et al.*, 1990). Deux régions ont été mises en évidence : celle allant du résidu 79 au 127 et celle allant du résidu 184 au 250 (figure 17). Une analyse de délétions de ce polypeptide indique qu'il est composé de deux domaines, un domaine N-terminal de liaison à l'autoinducteur et un domaine C-terminal de liaison à l'ADN (Choi and Greenberg, 1992a).

Le domaine N-terminal représente les deux tiers de LuxR. La phéromone, en se liant à ce domaine, va permettre de moduler l'activité du domaine C-terminal. D'autres analogues de l'autoinducteur peuvent se lier à LuxR. Certains servent d'inducteurs alternatifs de LuxR alors que d'autres interfèrent avec la fonction de la phéromone principale de LuxR. Les composés avec une substitution dans l'anneau de l'homosérine lactone ne se lient plus à LuxR (Schaefer *et al.*, 1996).

LuxR est associé au feuillet interne de la membrane cytoplasmique (Kolibachuk and Greenberg, 1993) mais nous ne connaissons pas la nature précise de cette interaction et sa signification. Il semble que le domaine N-terminal intervient dans cette interaction. Une hypothèse est que la liaison de la phéromone à LuxR est facilitée par des contacts protéine-lipides (Fuqua *et al.*, 1994). En absence d'autoinducteur, le domaine N-terminal bloque l'activation transcriptionnelle due au domaine C-terminal. Une hypothèse veut que ce soit causé par le blocage physique du domaine C-terminal par le domaine N-terminal qui empêcherait le premier de se lier à l'ADN (Stevens and Greenberg, 1999).

Certaines substitutions ponctuelles d'acides aminés dans LuxR résultent en une activation de la luminescence indépendamment de la présence d'autoinducteur. Ces substitutions touchent les deux domaines et cela pourrait signifier qu'il y a des interactions directes entre les deux domaines, influencées par l'autoinducteur (Sitnikov *et al.*, 1996) (cfr. figure 17).

Une fois que la phéromone appropriée est liée au domaine N-terminal, il y a probablement formation d'un dimère de LuxR (d'où la symétrie palindromique de la *lux* box) ; le domaine C-terminal peut alors activer la transcription. Une région comprise entre les résidus 116 et 160 est requise pour la dimérisation (Choi and Greenberg, 1992b). Un motif Hélice-Coude-Hélice (HTH) est situé entre les résidus 196 et 210 du domaine C-terminal et c'est ce motif qui est capital pour la liaison à l'ADN. D'autres études de délétants ont aussi montré que la région du domaine C-

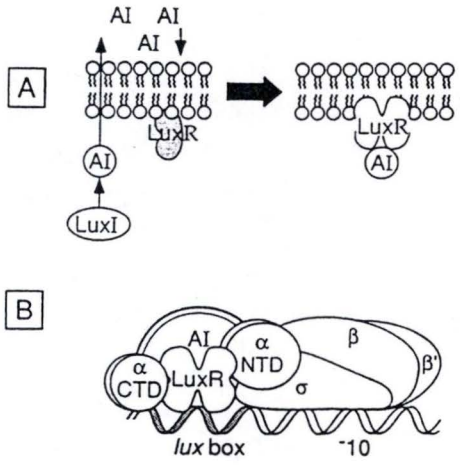


Figure 18 : Modèle du mécanisme de l'activation transcriptionnelle impliquant LuxR. (A) LuxI est responsable de la production d'autoinducteur. La membrane est perméable à cet autoinducteur qui s'accumule à haute concentration cellulaire. Il se lie au domaine N-terminal de LuxR associé à la membrane qui se dimérise alors et devient fonctionnel. (B) Un arrangement possible du complexe de transcription au promoteur *LuxI* (Stevens and Greenberg 1999).

terminal allant des résidus 230 à 250 est nécessaire pour l'activation de la transcription mais pas pour la liaison à l'ADN (Choi and Greenberg, 1992a) (cfr. figure 17).

2.2.2.3.L'étude biochimique de LuxR

Des études biochimiques de LuxR ont suivi les études génétiques de cette protéine, mais elles n'ont pas pu être menées *in vitro* sur la protéine LuxR prise en entier car quand elle est surexprimée dans une souche d'*E.coli* recombinante, toutes les LuxR traduites s'agrègent en un corps d'inclusion (Kaplan and Greenberg, 1987). L'association de LuxR avec la membrane cellulaire complique la purification. Il a cependant été possible de purifier une LuxR débarrassée de son domaine N-terminal. Puisque ce fragment purifié (LuxRΔN) active les gènes *lux* indépendamment de la présence d'autoinducteur, il peut être utilisé pour l'étude des interactions LuxR-ADN (Stevens *et al.*, 1994). Dans une expérience de protection à la DNaseI, LuxRΔN, plutôt que de protéger la *lux* box, protège une région non requise pour l'activation de la transcription (Stevens *et al.*, 1994). La sous-unité σ^{70} de la RNA polymérase (RNAP) protège, quand elle est seule, la région –10 du promoteur *luxI* mais pas la *lux* box ou la région –35 du promoteur *luxI*. Cependant, LuxRΔN et RNAP ensemble, protègent de manière coopérative la *lux* box et le promoteur *luxI* (Stevens *et al.*, 1994). Ceci suggère que les interactions protéine-protéine de ces deux fragments soient très importantes. Bien sûr, cette étude ne reflète peut-être pas la réalité puisque LuxRΔN est fort différente de LuxR.

La position de la *lux* box par rapport au promoteur *luxI* est la même que celle de la *CRP* box par rapport aux promoteurs *CRP* de classe II. Nous pouvons ainsi supposer que les interactions entre CRP et RNAP sont les mêmes que celles entre LuxR et RNAP (Ishihama, 1993) (figure 18). On pense que CRP interagit avec le domaine C-terminal de la sous-unité α de RNAP (α-CTD), avec le domaine N-terminal de cette même sous-unité (α-NTD) et avec la sous-unité σ (Busby and Ebright, 1997). L'interaction de l'α-CTD avec LuxR a été étudiée *in vivo* et *in vitro* et les résultats ont montré que l'α-CTD était essentielle pour l'activation de la transcription par LuxR. A part cela, on connaît peu de choses sur le mécanisme d'initiation de la transcription aux promoteurs de type *luxI* (Stevens and Greenberg, 1999).

Bien que le système LuxR/LuxI soit toujours le système de référence, des avancées récentes ont permis de récolter de plus amples informations sur le Quorum Sensing d'organismes tels qu' *Agrobacterium tumefaciens* et *Erwinia carotovora*. Il est très intéressant de posséder d'autres systèmes de référence car, comme nous l'avons vu ci-dessus, le système LuxR/LuxI présente certains inconvénients. Toutes ces nouvelles indications permettront de mieux comprendre le Quorum Sensing des espèces *Gram-négatives*. Ainsi, une étude récente a montré que, chez *E. carotovora*, la OHHL se lie à CarR avec une stoechiométrie de 2 moles de ligand pour 1 mole de

CarR sous forme dimérique. CarR semble préexister sous une forme dimérique à laquelle se lieraient deux phéromones, une par monomère. Ceci aurait pour effet d'augmenter la tendance de ce dimère à former un multimère et d'activer CarR. Une stoechiométrie similaire a pu être trouvée pour le complexe TraR – OOHL chez *A. tumefaciens*. Les données de cette étude montrent aussi que la capacité d'une HSL donnée à faciliter la liaison de CarR à sa cible d'ADN, est directement proportionnelle à l'affinité de l'HSL pour CarR (Welch *et al.*, 2000), (Zhu and Winans, 1999).

Malgré tous les progrès dans l'étude du Quorum Sensing, il n'existe toujours aucune structure cristallographique de protéines de type LuxR. Une telle structure peut pourtant donner de précieuses informations quant aux interactions entre la protéine de type LuxR et ses ligands (HSL, ADN), entre autres. Il est donc impératif de combler ce déficit de connaissances : parmi les différents outils disponibles, la bioinformatique est un moyen très intéressant pour aborder le sujet. En effet, cette approche théorique permet de contourner le manque de structures déterminées expérimentalement grâce à des méthodes de prédiction de structures protéiques de plus en plus performantes.

LA BIOINFORMATIQUE

3. Introduction

Depuis environ une dizaine d'années, les informations brutes générées par les biologistes moléculaires ont littéralement explosé. Les banques de données ont acquis une taille considérable et elles doublent de volume pratiquement chaque année. Cet accroissement impressionnant repose sur le rythme effréné auquel sont séquencés de nombreux génomes. Des génomes tels que celui de *E. coli*, *Saccharomyces cerevisiae, Arabidopsis thaliana* ont déjà été séquencés et on prévoit, pour dans quelques années (si pas quelques mois), le séquençage complet du génome humain. Il a donc été indispensable de fournir de nombreux programmes et moyens d'analyse pour pouvoir exploiter toutes ces données. Ainsi, la bioinformatique qui existe depuis que l'on archive des séquences d'acides d'aminés ou d'acides nucléiques, connaît pour la première fois une croissance très importante. Tout comme Darwin qui postula la théorie de l'évolution en comparant les caractéristiques morphologiques de plusieurs espèces animales, la comparaison des séquences de gènes et de protéines au moyen d'outils bioinformatiques peut révéler quelques précieuses informations (Baxevanis and Ouellette, 1998). C'est un domaine en évolution constante ; une mise à jour permanente des outils bioinformatiques et des travaux théoriques à la base de ce que l'on peut considérer comme une nouvelle science, permet aux biologistes d'utiliser aisément l'environnement de logiciels et de données. La quantité d'informations est trop grande pour être centralisée en un seul site. C'est pourquoi, le réseau Internet est le support idéal pour le développement de la bioinformatique et l'échange de données entre biologistes. Le grand problème, à l'heure actuelle, de cette accumulation très rapide d'informations dans les banques de données est qu'on ne peut assurer la fiabilité des données introduites. Ainsi, environ 5% des protéines contiennent des erreurs de séquençage. De plus, les annotations qui s'y rapportent sont basées sur l'homologie entre séquences similaires, or les protéines de référence ne sont pas toujours validées. A l'heure actuelle, le nombre d'outils bioinformatiques disponibles est énorme et la plupart de ceux-ci sont accessibles *via* le réseau.

EMBnet National Nodes		
Vienna Biocenter	Austria	http://www.at.embnet.org/
BEN	Belgium	http://www.be.embnet.org/
BioBase	Denmark	http://biobase.dk/
CSC	Finland	http://www.fi.embnet.org/
INFOBIOGEN	France	http://www.infobiogen.fr/
GENIUSnet	Germany	http://genome.dkfz-heidelberg.de/biounit/
IMBB	Greece	http://www.imbb.forth.gr/
HEN	Hungary	http://www.hu.embnet.org/
INCBI	Ireland	http://acer.gen.tcd.ie/
INN	Israel	http://dapsas.weizmann.ac.il/bcd/inn.html
IEN-ADR	Italy	http://bio-www.ba.cnr.it:8000/BioWWW/Bio-WWW.htm
CAOS/CAMM	Netherlands	http://www.caos.kun.nl/
Bio	Norway	http://www.no.embnet.org/
IBB	Poland	http://www.ibb.waw.pl/
PEN	Portugal	http://www.pen.gulbenkian.pt/
GeneBee	Russia	http://www.genebee.msu.su/
CNB-CSIC	Spain	http://www.es.embnet.org/
BMC	Sweden	http://www.embnet.se/
SIB	Switzerland	http://www.ch.embnet.org/
SEQNET	UK	http://www.seqnet.dl.ac.uk/
EMBnet Specialist Nodes		
MIPS	Germany	http://www.mips.biochem.mpg.de/
ICGEB	Italy	http://www.icgeb.trieste.it/
Pharmacia Upjohn	Sweden	http://www.pnu.com/
F.Hoffmann-La Roche	Switzerland	http://www.roche.com/
EBI	UK	http://www.ebi.ac.uk/
HGMP-RC	UK	http://www.hgmp.mrc.ac.uk/
Sanger	UK	http://www.sanger.ac.uk/
UCL	UK	http://www.biochem.ucl.ac.uk/bsm/dbbrowser/embnet/
EMBnet Associate Nodes		
IBBM	Argentina	http://sol.biol.unlp.edu.ar/
ANGIS	Australia	http://www.angis.su.oz.au/
CBI	China	http://www.cbi.pku.edu.cn/
CIGB	Cuba	http://bio.cigb.edu.cu/
CDFD	India	Not yet available
SANBI	South Africa	http://www.sanbi.ac.za
USA Information Providers		
NCBI	USA	http://www.ncbi.nlm.nih.gov/
NLM	USA	http://www.nlm.nih.gov/
NIH	USA	http://www.nih.gov/

Tableau 3 : Fournisseurs européens et américains d'informations bioinformatiques (Attwood and Parry-Smith 1999)

4. Les outils bioinformatiques

4.1. Les banques de données

Au milieu des années 80, les banques biologiques ont commencé à proliférer et la demande de moyens pour accéder à ces données se faisait de plus en plus pressante. Les organisations scientifiques mondiales ont entrevu l'immense potentiel du réseau Internet pour la communication et la centralisation des ressources. En 1988, un réseau a été établi pour relier les laboratoires européens qui se servaient de la bioinformatique. Il a été appelé EMBnet. Il s'est fortement développé depuis 12 ans et regroupe actuellement de nombreux fournisseurs d'informations bioinformatiques. L'étape suivante de la création d'EMBnet a été de faciliter la consultation de toutes ces données regroupées en un site particulier, à l'aide d'une interface aisée : un projet de recherche a abouti à la création du « Sequence Retrieval System » (SRS). En Amérique, le fournisseur d'informations est le NCBI (« National Center for Biotechnology Information ») qui a été établi en 1988 comme une division de la NLM (« National Library of Medicine ») (tableau 3). L'équivalent américain du SRS d'EMBnet est le système Entrez de la NCBI qui fournit l'accès à des banques de séquences d'ADN, de protéines, de structures de protéines, à des données d'annotations de chromosomes et de génomes et à la banque de données bibliographiques PubMed (Attwood and Parry-Smith, 1999) (tableau 4).

4.1.1.Les banques de séquences nucléiques

Les principales sont Genbank (USA), EMBL (Europe) et DDBJ (Japon) qui échangent journalièrement leurs nouvelles données introduites (tableau 4). EMBL qui est la banque de séquences d'ADN de l'EBI[1] (« European Bioinformatics Institute »), contient plus de 6 millions d'entrées. La DDBJ du NIG (« National Institute of Genetics ») commença sa collaboration avec les deux autres banques en 1986. Genbank qui est installée au NCBI est divisée en 17 sections pour faciliter la recherche spécifique en la limitant à des sections particulières telles que PRI (primates), EST (« Expressed Sequence Tags »), BCT (bactéries), INV (invertébrés),...(Attwood and Parry-Smith, 1999).

4.1.2.Les banques de séquences protéiques

PIR, SWISS-PROT et TrEMBL sont les plus connues (tableau 4). PIR qui appartient à la NBRF (« National Biomedical Research Foundation ») est divisée en 4 sections qui diffèrent les unes des autres par la qualité des données et le nombre

[1] Il s'agit d'une branche de l'EMBnet

Tableau 4 :

PubMed	http://www.ncbi.nlm.nih.gov/PubMed/
SRS	http://srs.ebi.ac.uk/
Entrez	http://www.ncbi.nlm.nih.gov/Entrez/
GenBank	http://www.ncbi.nlm.nih.gov/Web/Genbank/
EMBL	http://www.ebi.ac.uk/ebi_docs/embl_db/ebi/topembl.html
DDBJ	http://www.ddbj.nig.ac.jp/
PIR	http://nbrfa.georgetown.edu/pir/
SWISS-PROT	http://expasy.hcuge.ch/sprot/sprot-top.html
PROSITE	http://expasy.hcuge.ch/sprot/prosite.html
BLOCKS	http://www.blocks.fhcrc.org/
SCOP	http://scop.mrc-lmb.cam.ac.uk/scop/
CATH	http://www.biochem.ucl.ac.uk/bsm/cath/
PDB	http://pdb-browsers.ebi.ac.uk//index.html
PSORT	http://psort.nibb.ac.jp/
TMpred	http://www.ch.embnet.org/software/TMPRED_form.html

d'annotations fournies. SWISS-PROT a été produite en 1986 par le Département de Biochimie Médicale de l'Université de Genève et par l'EMBL. Elle est actuellement entretenue par la SIB (« Swiss Institute of Bioinformatics ») et par l'EBI. En 1996, un supplément de SWISS-PROT a été créé. Ce supplément est la banque TrEMBL. Elle permet de conserver des séquences moins annotées que celles de SWISS-PROT.

Il est intéressant de consulter les trois banques citées ci-dessus car elles fournissent chacune des informations complémentaires. Avec trois recherches, on trouvera cependant beaucoup de redondances. Il existe ainsi des banques telles que la NRDB (« Non-Redundant DataBase ») qui amalgament les informations de différentes banques. La NRDB est une combinaison de GenPept (dérivée de Genbank), de PDB, de SWISS-PROT, de PIR et de GenPeptupdate (mise à jour de GenPept). Les séquences ne s'y retrouvent pas deux fois mais il persiste cependant quelques problèmes ; par exemple, certaines protéines y sont introduites plusieurs fois avec des erreurs de séquençage différentes. Cette banque est la banque par défaut pour une recherche avec BLAST (Attwood and Parry-Smith, 1999).

4.1.3.Les banques de données secondaires

Elles sont appelées ainsi car elles résultent de l'analyse de séquences des banques de données primaires. SWISS-PROT est la banque de données primaire la plus utilisée par les banques de données secondaires. Celles-ci sont utiles car elles contiennent des informations sur les motifs qui sont des régions très conservées entre protéines homologues. Les motifs jouent d'habitude un rôle biologique vital et spécifique. Une séquence inconnue peut être confrontée à une telle banque pour déterminer si elle contient un motif qui pourrait la rattacher à une famille de protéines aux fonctions connues. La première banque de données secondaires est PROSITE qui est maintenue à la SIB (tableau 4). Les motifs de cette banque dérivent de la construction d'alignements de plusieurs séquences homologues et de l'inspection manuelle des régions conservées ; les motifs représentent alors un consensus de ces régions. Un autre exemple est BLOCKS (tableau 4). Les motifs ou blocs sont créés automatiquement en détectant les régions les plus conservées de chaque famille de protéines inclues dans la banque PROSITE. Les blocs sont encodés en tant qu'alignements multiples sans « gaps » (cfr. point 2.4.2) (Attwood and Parry-Smith, 1999).

4.1.4.La banque PDB

Les méthodes expérimentales de détermination de structures, c'est-à-dire la diffraction aux rayons X et la NMR, permettent la résolution d'un grand nombre de structures. Le stockage de celles-ci dans une banque de structures est donc essentiel. Cette banque est la « Protein Data Bank » (PDB) (tableau 4). Elle contient les coordonnées cartésiennes de la structure de protéines, d'acides nucléiques et d'hydrates de carbones. La PDB n'est pas exempte de tout défaut ; ainsi, le nombre

de protéines originales qui s'y trouvent est très largement inférieur au nombre d'entrées. En effet, de nombreuses protéines se retrouvent en plusieurs exemplaires sous forme de complexes avec divers inhibiteurs ou alors avec des résidus mutés. La redondance est donc importante. Il existe une extension à cette banque nommée « Pending/Waiting List ». Elle regroupe les protéines dont la structure est en cours de détermination. La présence d'une protéine dans la « Pending/Waiting List » indique la disponibilité proche de sa structure 3D, nouvellement résolue.

4.1.5.Les banques de classification de structures

De nombreuses protéines partagent une similarité structurale reflétant dans certains cas une origine évolutive commune. Pour pouvoir détecter plus facilement les caractères évolutifs communs entre protéines, différents types de classification de structures ont été établis. Les deux plus connus sont SCOP (« Structural Classification of Proteins ») maintenu au « MRC Laboratory of Molecular Biology » et CATH (« Class, Architecture, Topology, Homology ») maintenu à la « University College London » (UCL) (tableau 4). SCOP classe les structures de protéines d'une façon hiérarchique. Les différents niveaux sont la classe, le « repliement », la superfamille, la famille et l'espèce. CATH utilise cinq niveaux hiérarchiques qui sont la classe, l'architecture, la topologie, l'homologie (la superfamille) et la séquence (la famille) (Attwood and Parry-Smith, 1999).

4.2. Outils de prédictions de localisation

Il existe de nombreux programmes disponibles sur le réseau qui permettent de tirer toutes sortes d'informations à partir de la séquence. On peut citer, par exemple, PSORT ou encore TMpred (tableau 4), la liste étant bien sûr beaucoup plus longue. PSORT est un programme qui prédit la localisation de protéines dans des cellules (Nakai and Kanehisa, 1991;Nakai and Horton, 1999). Il se base sur la séquence de la protéine d'intérêt et sur son origine (bactérie *Gram-positive* ou *Gram-négative*, cellule animale, levure,...). Selon la présence d'un peptide signal, d'un segment trans-membranaire, d'une région typique aux lipoprotéines (aux alentours du site de clivage),... PSORT pourra prédire la localisation subcellulaire de la protéine. TMpred, lui, prédit les régions protéiques trans-membranaires et leur orientation. L'algorithme est basé sur la comparaison de l'analyse statistique de Tmbase, qui est une banque de protéines trans-membranaires et de leur domaine trans-membranaire (Hofmann and Stoffel, 1993), avec la séquence d'intérêt pour en retirer la prédiction.

4.3. La prédiction de structures secondaires

La prédiction de structures secondaires est l'étape la plus générale pour obtenir des informations structurales sur une nouvelle séquence. Cette prédiction peut, par exemple, aider à confirmer des relations structurales et fonctionnelles entre des protéines peu similaires au niveau de leur séquence. Elle peut aussi être importante dans l'établissement d'alignements. En effet, les structures secondaires apportent de l'information supplémentaire pour aligner plusieurs séquences. Les méthodes de prédictions de structures secondaires sont divisées en trois groupes (Sternberg, 1996) :

- les méthodes statistiques :

Elles sont basées sur l'étude de protéines de structures primaires et secondaires connues. Le principe est de rechercher des relations empiriques entre les deux types de structures. Les premières méthodes souffraient du manque de données dans les banques. Aujourd'hui, les banques sont plus fournies mais le nombre de structures disponibles reste toujours un facteur limitant. Une méthode qui fut populaire en son temps est celle de Chou et Fasman (Chou and Fasman, 1974). Cependant, on a aujourd'hui largement démontré le caractère aléatoire des résultats qu'elle fournit ; elle doit donc être évitée. D'autres méthodes plus confiantes existent : celle de GOR (Garnier *et al.*, 1978) ou encore celles qui sont basées sur les réseaux neuronaux telles que PHD (Rost and Sander, 1993) (tableau 6).

- les méthodes physico-chimiques :

Elles sont basées sur les connaissances des bases physico-chimiques des structures protéiques, par exemple, la méthode de Lim (Lim, 1974).

- les méthodes hybrides :

Elles combinent les meilleurs aspects des deux types de méthodes précédents. Les programmes utilisant ces algorithmes capables d'apprendre des règles de prédiction sont, par exemple, PROMIS (King and Sternberg, 1990) ou GOLEM (Muggleton *et al.*, 1992).

Les méthodes antérieures à PHD donnent en général de mauvais résultats (inférieurs à 60% de prédictions exactes). Aujourd'hui, PHD n'est plus la meilleure méthode (cfr matériel et méthodes). En général, si nous possédons uniquement la séquence primaire, les meilleures méthodes prédiront les structures secondaires avec 65% d'exactitude ; si un certain nombre d'homologues sont connus, ce chiffre monte à 70% d'exactitude; et si le domaine structural de la protéine est connu, l'exactitude peut monter jusqu'à 80% pour certains domaines. Il est recommandé tout de même, d'utiliser plusieurs approches et de combiner les résultats en un alignement, pour en

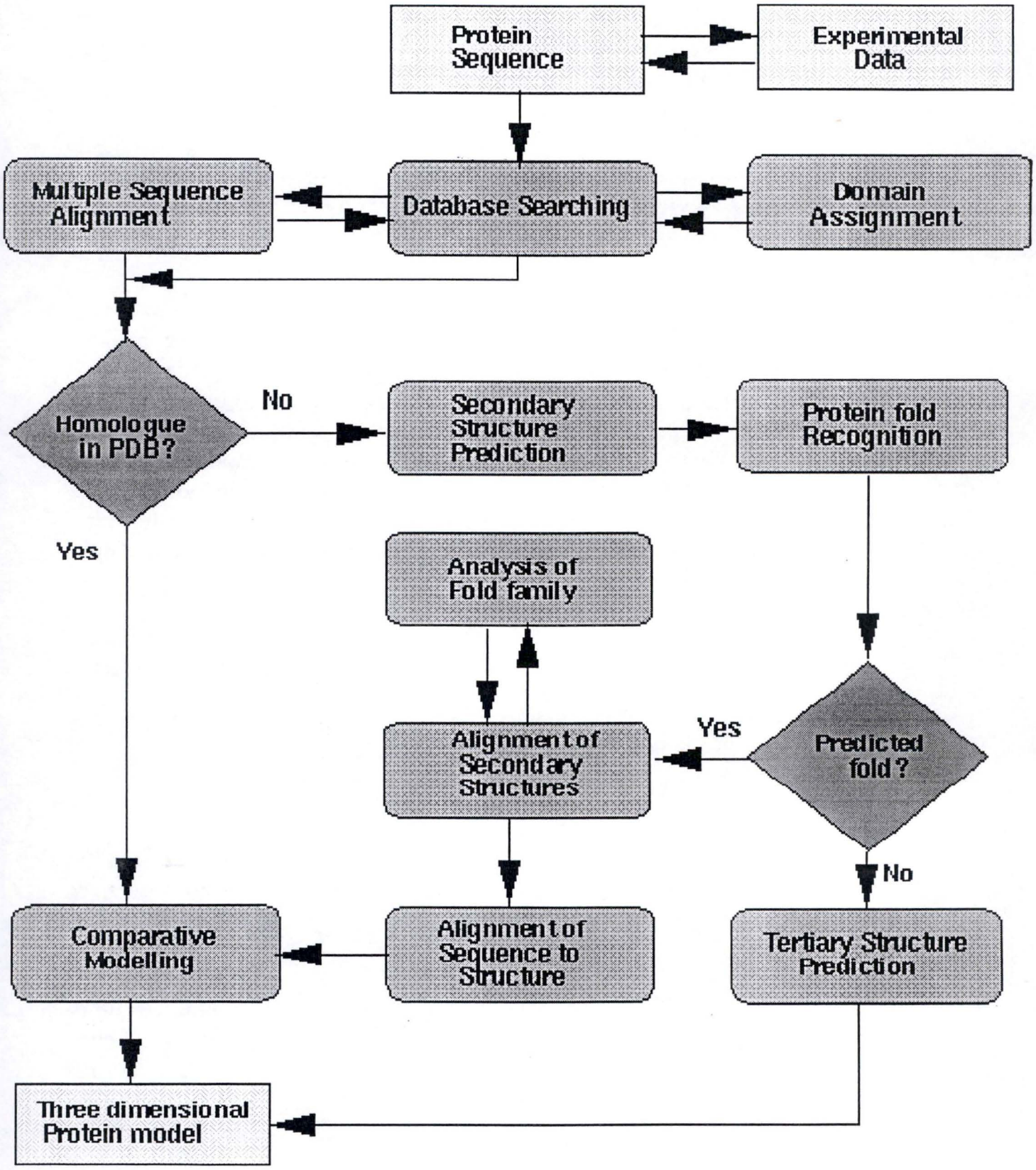


Tableau 5 : Diagramme des différentes étapes parcourues lors de prédictions de structures de protéines (Russel 1999)

retirer une prédiction « consensus » beaucoup plus précise et plus fiable (King *et al.*, 2000).

4.4. La détermination de la structure tridimensionnelle d'une protéine

Connaître la structure tridimensionnelle d'une protéine présente un grand intérêt scientifique pour de multiples raisons. Elle aide à la compréhension des caractéristiques protéiques telles que les interactions entre sous-unités protéiques, les interactions avec un ligand ou avec l'ADN, l'activité enzymatique au site actif ou encore la stabilité. Il est actuellement reconnu que la fonction d'une protéine est fortement conditionnée par sa structure.

Plusieurs méthodes ont été élaborées pour fournir un modèle tridimensionnel d'une protéine d'intérêt. D'un côté, il y a les méthodes expérimentales incluant la diffraction aux rayons X et la Résonance Magnétique Nucléaire (RMN) ; de l'autre, il y a les méthodes théoriques bioinformatiques dont le principe est de prédire la structure à partir de la séquence. Les premières présentent certains inconvénients malgré leur grande précision : la diffraction aux rayons X requiert au préalable des cristaux protéiques dont l'obtention est hasardeuse et la RMN ne donne pas toujours de résultats ou des résultats parfois difficilement interprétables. Les méthodes théoriques, quant à elles, donnent des modèles tridimensionnels relativement moins confiants.

La prédiction de structures protéiques tridimensionnelles est devenu un problème pressant pour les biologistes[1] car le fossé entre le nombre de structures déterminées expérimentalement et le nombre de séquences disponibles dans les banques de données ne cesse de s'agrandir (Sternberg, 1996). En effet, il y a actuellement plus de 490 000 entrées dans la banque de séquences protéiques non-redondantes (NR) mais seulement environ 12 000 dans la banque de structures 3D, PDB (« Protein Data Bank » Brookhaven National Laboratory -Cambridge USA). Cependant, la modélisation ne présente pas un intérêt réel pour absolument toutes les protéines étudiées ; cela dépend en effet du type de données que l'on recherche sur la protéine. Les tentatives de combler ce vide reposent sur les méthodes prédictives bioinformatiques. Ces méthodes peuvent fournir certaines caractéristiques d'une protéine même sans données biochimiques. C'est pourquoi, le nombre de méthodes élaborées à cette fin a beaucoup augmenté ces dernières années et leur puissance augmente continuellement. Les différentes étapes générales de modélisation sont reprises dans le tableau 5.

[1] Le problème a également été soulevé au CASP (« Critical Assessment of methods of Protein Structure Prediction »). Ce dernier est un « concours » se déroulant tous les deux ans (cette année se déroule le CASP4). Les participants appliquent leur méthode de prédictions de structure à des séquences dont la structure tridimensionnelle est bientôt disponible mais auxquelles ils n'ont aucun accès. Après cela, le modèle de chaque participant est comparé à la structure réelle. Les résultats sont publiés à chaque fois dans un supplément de Proteins. Le CASP permet d'évaluer les méthodes et est donc considéré comme une référence par les bioinformaticiens se lançant dans la prédiction de structures.

Tableau 6 :

BLAST	http://www.ncbi.nlm.nih.gov/blast/blast.cgi
PSI-BLAST	http://www.ncbi.nlm.nih.gov/blast/psiblast.cgi
FASTA	http://www2.ebi.ac.uk/fasta3/
ALIGN	http://vega.crbm.cnrs-mop.fr/bin/align-guess.cgi
MATCHBOX	http://www.fundp.ac.be/sciences/biologie/bms/matchbox_submit.html
ClustalW	http://dot.imgen.bcm.tmc.edu:9331/multi-align/Options/clustalw.html
SWISSMODEL	http://www.expasy.ch/swissmod/SWISS-MODEL.html
PHD	http://www.embl-heidelberg.de/Services/sander/predictprotein/
3DPSSM	http://www.bmm.icnet.uk/~3dpssm/
Topits	http://www.embl-heidelberg.de/Services/sander/predictprotein/

4.4.1.Recherche de similarité dans les banques de données

La première étape qui suit l'acquisition d'une nouvelle séquence consiste en la recherche dans les banques de données, de séquences similaires à la séquence cible[1]. Cela nous permet de déterminer quelles sont, parmi les centaines de milliers de séquences disponibles dans les banques, celles qui sont potentiellement en relation avec la séquence d'intérêt. Il est intéressant de retrouver dans les résultats au moins une protéine dont on connaît la structure 3D. Elle servira de séquence de référence (ou séquence « template ») pour l'élaboration du modèle 3D au moyen de la modélisation par homologie. Si ce n'est pas le cas, il faudra emprunter une autre voie (cfr. point 2.4.4).

Dans ces recherches en banques de données, BLAST suit ces différentes étapes :

1. recherche de courts « mots » conservés (HSPs, « High score Segment Pairs ») dans la banque de séquences
2. tentative d'étendre la similarité de part et d'autre des HSPs
3. calcul d'un score quantifiant la similarité et d'une « Expected-value » (nombre de séquences ayant obtenu un score donné, qu'on attendrait au hasard, dans une banque de taille donnée, avec une séquence cible de taille déterminée)

Ces recherches sont exécutées aisément grâce à des programmes accessibles sur le réseau Internet. Les plus connus sont BLAST (Altschul *et al.*, 1990) et FASTA (Pearson and Lipman, 1988). PSI-BLAST, une version améliorée de BLAST, fait aussi ses preuves (Altschul *et al.*, 1997) (tableau 6). C'est une méthode itérative ; elle commence par une recherche standard dans une banque de données et un profil des alignements réalisés par cette recherche initiale est construit. Le profil est une matrice $20 \times n$, n étant le nombre de colonnes dans l'alignement ; chaque case donne la fréquence de l'acide aminé correspondant à la position correspondante dans l'alignement. Ce profil est utilisé dans une seconde recherche. Ce processus peut être répété pour trouver de nouvelles séquences lors de chaque cycle et affiner ainsi le profil. PSI-BLAST est donc capable de retrouver des similarités faibles mais biologiquement significatives entre des séquences.

Il est recommandé d'utiliser plusieurs banques de séquences telles que SWISS-PROT, PIR ou GENBANK afin de disposer du plus grand nombre de séquences similaires possibles pour augmenter la fiabilité de l'alignement que l'on obtiendra dans l'étape suivante. Une banque intéressante est la banque NR qui regroupe toutes les séquences non-redondantes déterminées à ce jour. Il est important aussi de faire ces recherches dans une banque de séquences pour laquelle on possède des informations sur la structure 3D, c'est-à-dire dans la banque PDB. En effet, ces

[1]La séquence cible ou séquence « target » est la protéine pour laquelle on cherche les caractéristiques de sa structure 3D. On parlera de séquence « template » pour la protéine de structure 3D connue que l' on utilise comme référence pour modéliser la séquence cible.

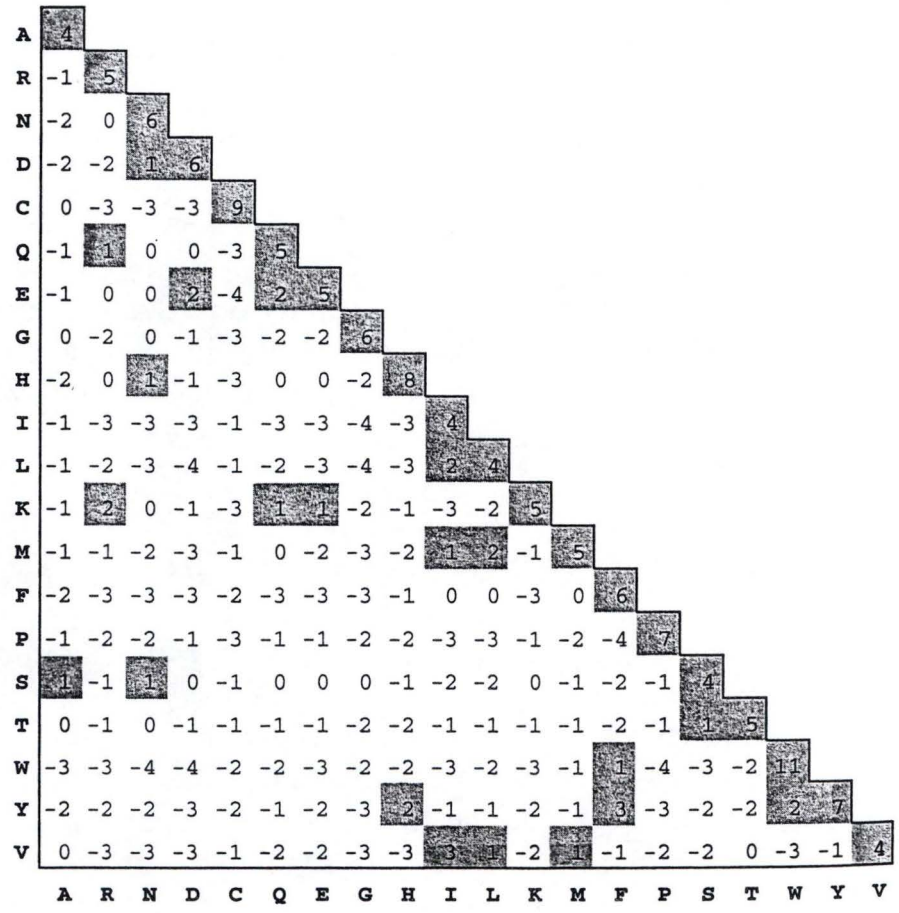


Figure 19 : La matrice de score BLOSUM 62 (Baxevanis and Ouellette 1998)

informations n'apparaissent pas clairement dans la recherche de séquences à travers d'autres banques de données, car elles sont « noyées » dans la masse de séquences similaires (Russel, 1999).

4.4.2.L'alignement de séquences

Dans le contexte de la prédiction de structures de protéines, il y a plusieurs raisons pour réaliser un alignement de séquences protéiques :

A . Pour déterminer les régions structuralement conservées[1] entre la séquence cible et la séquence de structure connue qui servira de référence

B . Pour identifier la fonction potentielle de la séquence nouvellement déterminée, par analogie avec des protéines de fonction connue

C . Pour attribuer sur la séquence cible, une fonction aux résidus ou aux régions conservés, par analogie avec une protéine bien caractérisée ; en effet, ces résidus ou régions sont susceptibles de garder leur fonction d'une protéine/d'un domaine homologue à l'autre

D . Pour apporter une information sur des résidus supposés enfouis ou exposés au solvant ; en effet, les régions hydrophobes/hydrophiles conservées dans l'alignement, seront plus probablement à l'intérieur de la protéine/ exposées au solvant

E . Les alignements fournis en « input » aux méthodes de prédictions de structures augmentent la fiabilité de celles-ci.

Tous les algorithmes utilisés pour comparer des séquences protéiques reposent sur un système de scores caractérisant la similarité ou la distance entre chaque paire d'acides aminés possible c'est-à-dire 210 paires[2]. Ces scores sont rassemblés dans une matrice 20x20. Il existe plusieurs types de matrice telle que la matrice de Dayhoff (Dayhoff *et al.*, 1972), la matrice BLOSUM (Henikoff and Henikoff, 1992), ou encore la matrice d'Overington (Overington *et al.*, 1990), utilisant chacune un système de scores différent (figure 19).

Une fois la matrice de score choisie, le problème suivant est de générer un alignement. La question se subdivise en différentes méthodes existant pour les alignements pairés (de deux séquences), les alignements multiples (de plusieurs séquences) et les méthodes incorporant des informations telles que les structures secondaires ou tertiaires des protéines. Les méthodes les plus simples sont celles qui n'introduisent pas d'insertions ou de délétions (« gaps »). Ces « gaps » permettent, par exemple, d'aligner plus efficacement des séquences de longueurs différentes.

---

[1] Des séquences ou des régions homologues ont une grande chance d'adopter la même structure.
[2] 190 paires d'acides aminés différents et 20 paires d'acides aminés identiques

Des programmes tels que FASTA (Pearson and Lipman, 1988) ou encore BLAST (Altschul *et al.*, 1990) utilisent des algorithmes d'alignement pairé pour des recherches en banques de données. D'autres tels que ALIGN sont uniquement destinés à aligner deux séquences fournies (Myers and Miller, 1988) (tableau 6).

Les alignements multiples sont plus fiables que les alignements pairés. En effet, le risque de faux positifs est amoindri : une région similaire dans plusieurs séquences a beaucoup plus de poids qu'une région similaire partagée par seulement deux séquences car le hasard ne sera pas à l'origine de cette similarité. Certains programmes sont spécifiques à l'alignement de plusieurs séquences, comme Match-Box (Depiereux and Feytmans, 1992) & (Depiereux *et al.*, 1997), ou ClustalW (Thompson *et al.*, 1994) (tableau 6).

4.4.3.Modélisation par homologie

La prédiction de la structure 3D d'une protéine est beaucoup plus précise si la structure d'un ou plusieurs homologues est connue. Les informations structurales de ce(s) dernier(s) peuvent alors être extrapolées à la nouvelle séquence. En effet, deux séquences de plus de 70 acides aminés partageant au moins 30% d'identité ont une très forte probabilité d'adopter la même conformation générale (Doolittle, 1981). Le modèle obtenu peut servir de point de départ à la détermination structurale d'une protéine par RMN ou par diffraction des RX. Plusieurs étapes constituent ce procédé de modélisation (Vinals, 1996) :

- la détermination de régions structuralement conservées (SCRs) sur base d'alignements multiples de séquences et éventuellement de structures

- l'établissement des correspondances entre la séquence cible et la (les) structure(s) similaire(s) connue(s) à partir de l'alignement multiple

- la construction d'une partie du squelette sur base des SCRs

- le positionnement des chaînes latérales

- l'optimisation par mécanique moléculaire et/ou dynamique moléculaire pour corriger des invraisemblances énergétiques introduites dans le modèle. Ces invraisemblances peuvent être des liaisons covalentes inadéquates, des contacts stériques défavorables,...

La dernière étape de ce processus exige de connaître les différentes énergies associées aux atomes et à leurs interactions. Une fonction d'énergie doit donc être décrite. La minimisation d'énergie repose sur le principe de faire évoluer le système vers la conformation d'énergie minimale sans garantir que ce minimum soit le

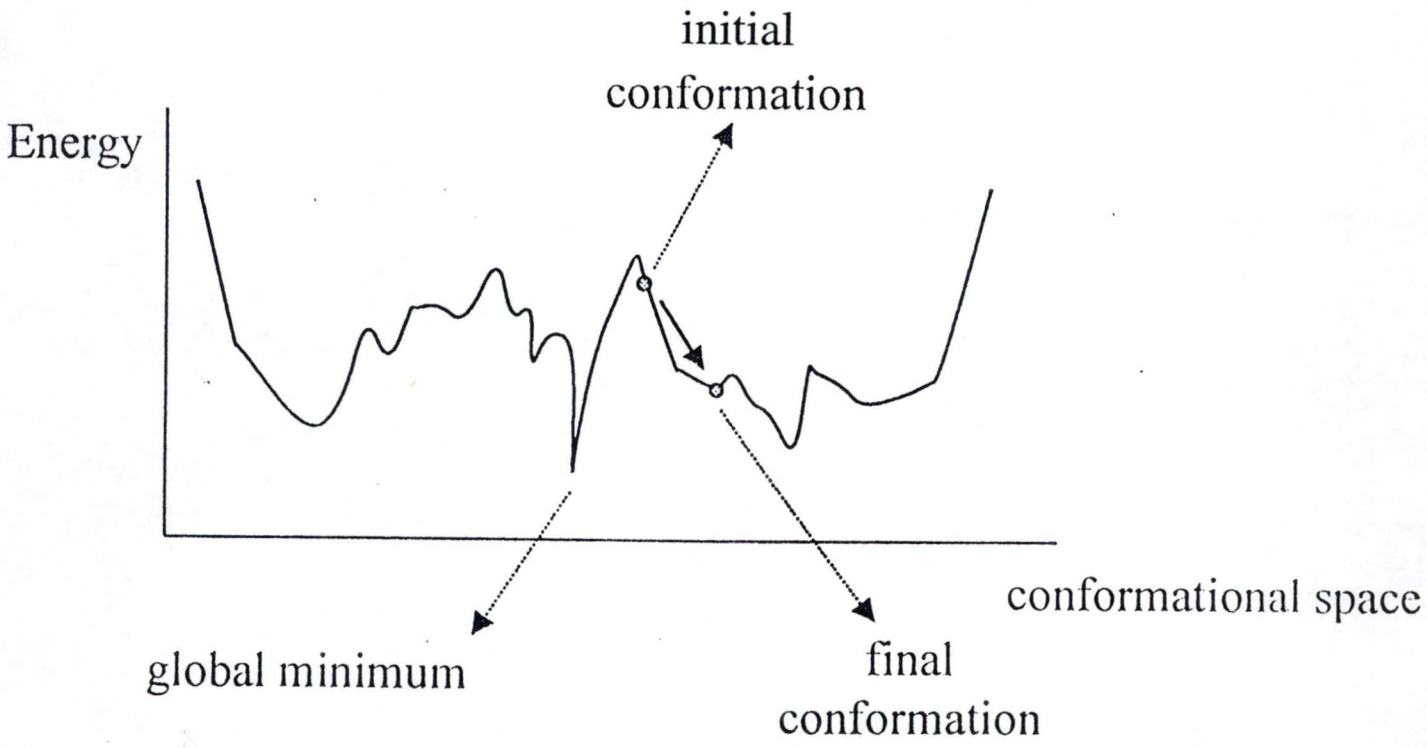


Figure 20 : La conformation initiale est séparée du minimum d'énergie global par une barrière énergétique. La minimisation d'énergie aboutira au minimum d'énergie local.

minimum global de la fonction d'énergie[1]. Si ce minimum global est séparé de l'état initial par une barrière énergétique, un algorithme de minimisation d'énergie est incapable de l'atteindre (figure 20). La dynamique moléculaire permet au système de passer ces barrières énergétiques grâce à de l'énergie cinétique additionnelle et d'aboutir à un minimum d'énergie local potentiellement plus proche du minimum d'énergie global.

- l'évaluation qui permet de vérifier les structures en utilisant, par exemple, un potentiel de force moyenne

SWISSMODEL est un serveur de modélisation par homologie automatisée qui consiste en une approche à deux étapes (Guex and Peitsch, 1997) (tableau 6). Premièrement, on envoie une séquence vers SWISSMODEL qui va la comparer aux séquences d'ExPDB (banque dérivée de PDB). Le processus se poursuit uniquement si un ou plusieurs homologues sont retrouvés dans ExPDB. Deuxièmement, un modèle atomique est construit et retourné par courrier électronique.

4.4.4.Modélisation de protéines sans homologue de structure 3D connue

4.4.4.1.Reconnaissances de « folds » (repliements)

Des méthodes de prédictions de structures tertiaires basées sur la reconnaissance de « folds » appropriés à partir d'une séquence d'acides aminés sont apparues de l'observation suivante : deux structures peuvent avoir le même repliement malgré une absence statistiquement significative de similarité de séquence. Bien que certaines de ces paires de protéines possèdent des fonctions similaires, la plupart d'entre-elles ne montrent aucune ressemblance fonctionnelle. Cela suggère qu'il n'existerait qu'un nombre limité de topologies ou « repliements » sélectionnés par l'évolution. La question est alors de se demander si une séquence donnée pourrait adopter un « repliement » actuellement repris dans les banques de structures 3D. Contrairement à la seule comparaison de séquences, ces méthodes profitent des informations sur les structures 3D connues. En effet, le problème est pris à l'envers puisque, plutôt que de prédire la structure tertiaire à partir de la séquence, ces méthodes essayent de déterminer comment un « repliement » convient à la séquence protéique. Elles sont classées en deux catégories :

- les méthodes de THREADING (= action d'enfiler) :

Ces méthodes sont basées sur des algorithmes qui « enfilent » la séquence protéique sur une structure tridimensionnelle connue et qui déterminent un

[1] Il correspond au minimum absolu de la fonction d'énergie libre, somme du potentiel intramoléculaire, de l'entropie et de l'énergie libre de solvatation.

alignement correspondant à un modèle structural énergétiquement favorable. Un calcul d'énergie est réalisé pour chaque « enfilement » sur tous les « repliements » connus et ce sont les « repliements » produisant les énergies les plus favorables qui sont sélectionnés dans les résultats. Le potentiel énergétique étant pris en compte, ces méthodes peuvent prédire une structure tertiaire même s'il n'y a pas de similarité. Elles incluent des méthodes telles que THREADER (Miller *et al.*, 1996) ou ProFIT (Sippl, 1993). Ce qu'il faut remarquer, c'est que l'énergie à laquelle nous faisons référence ici n'est pas une énergie estimée par un champ de force. La mesure de l'énergie utilisée pour le « threading » est un potentiel de force moyenne dérivé, par une loi de Boltzmann inverse, de la fréquence de rencontre de résidus dans une conformation particulière.

- les méthodes de Pseudo-THREADING :

Ces méthodes ne se basent pas sur des calculs de potentiels énergétiques empiriques mais plutôt sur la comparaison de structures secondaires et de patterns d'accessibilité entre la séquence cible et la protéine de « repliement » connu. Elles alignent ensuite les motifs similaires. La séquence est finalement « enfilée » sur les « repliements » pour lesquels les alignements de ces motifs sont les meilleurs. On obtient ainsi un alignement de la séquence à la structure tertiaire. Des méthodes telles que Topits (Rost, 1995) utilisent le principe des réseaux neuronaux. Une autre méthode que nous pouvons citer est 3DPSSM qui s'est avérée être intéressante au CAFASP-1 (« Critical Assessment of Fully Automated Structure Prediction Methods ») (Fischer *et al.*, 1999) (tableau 6).

Ces deux classes de méthodes ne sont pas à la pointe de la perfection. Nous avons encore besoin de meilleurs potentiels, de meilleures méthodes, pour générer les structures secondaires et de meilleures techniques pour obtenir le threading optimal (Sternberg, 1996). Nous avons une chance sur deux seulement d'avoir le « repliement » correct situé en tête de la liste de résultats. Ces méthodes déterminent un alignement séquence-structure mais certaines d'entre-elles introduisent des « gaps » d'une façon critiquable. Ainsi, même si elles donnent le bon « repliement » en tête de liste, le modèle pourrait être faux dû au mauvais alignement séquence-structure.

Une fois les étapes précédentes réalisées, on possède assez d'informations pour concevoir l'alignement séquence-structure 3D optimal qui permettra de modéliser la séquence cible. Il faut s'assurer que plusieurs critères sont respectés :

- l'alignement des résidus prédits comme enfouis ou exposés au solvant avec ceux connus être enfouis ou exposés dans la structure « template »

- la conservation plus ou moins grande des propriétés des résidus de la structure connue (taille, hydrophobicité, polarité,...) pour la séquence cible

- la non-disruption des ponts H dans des structures composées de plans β. Ceci se fait grâce à un pseudo-potentiel mesurant les interactions résidus-résidus et les interactions entre brins β pour n'importe quelle protéine (Hubbard *et al.*, 1996).

4.4.4.2. Méthodes de prédictions *ab initio*

Comme son nom l'indique, ce type de prédiction se fait sans aucune donnée de départ, excepté la structure primaire de la protéine cible. La mécanique et la dynamique moléculaire interviennent pour calculer une conformation d'énergie minimale. Elles prédisent une conformation au niveau atomique bien que celle-ci puisse contenir des erreurs. Cette grande précision requiert cependant un nombre incroyablement élevé de calculs. A ce jour, aucune méthode n'a pu donner des résultats convaincants pour des molécules plus importantes que des peptides de 40 à 60 acides aminés. La puissance actuelle des ordinateurs limite aussi ces résultats. Les méthodes de ce type sont, par exemple, Monte Carlo (Holm and Sander, 1992), la méthode de Brasseur (Brasseur, 1995), ou encore celle développée par Gilis et Rooman (Gilis, 1999),(Gilis and Rooman, 2000).

4.5. Le « docking » de biomolécules

Les interactions entre biomolécules jouent un rôle majeur dans les mécanismes biologiques. Elles permettent entre autres de maintenir les réseaux complexes d'interactions métaboliques et régulatoires. Leur analyse requiert la connaissance d'un certain nombre de caractéristiques de la structure 3D des molécules impliquées. Les méthodes bioinformatiques de « docking » visent, soit une analyse détaillée d'interactions entre biomolécules, soit une détermination du ligand correspondant à une protéine d'intérêt par la recherche au travers de banques de ligands. Ces études permettraient, par exemple, de découvrir ou de synthétiser des inhibiteurs très affins qui se lieraient mieux à la protéine que le ligand naturel ; dans le cas de protéines appartenant à des bactéries pathogènes, elles pourraient être inhibées ce qui pourrait supprimer ou diminuer la virulence.

La plupart des analyses bioinformatiques de « docking » moléculaire impliquent au moins une protéine comme partenaire. Les protéines peuvent se lier à l'ADN, à l'ARN, à d'autres protéines ou encore à de petits composés organiques ou métalliques. Suivant le type de partenaire de la protéine, le problème du « docking » requiert différents algorithmes (Sternberg *et al.*, 1998 pour une revue), (Lengauer and Rarey, 1996 pour une revue). Le « docking » ne se fait pas de façon triviale ; plusieurs difficultés peuvent apparaître. Ainsi, par exemple, il faut pouvoir trouver le positionnement de départ des partenaires même s'il n'est qu'approximatif. De plus, certaines erreurs peuvent être commises pour de grandes molécules. Les petites molécules posent moins de problèmes et des méthodes automatisées ont d'ailleurs été

développées pour elles, telles que le programme AutoDock (Goodsell and Olson, 1990), (Morris *et al.*, 1998).

4.5.1.Une protéine comme partenaire

La plupart des approches de « docking » protéine-protéine sont basées sur l'hypothèse du « corps rigide ». Ce modèle très simplifié considère les deux protéines comme des corps solides rigides. Des modèles de surfaces géométriques et des données sur les structures 3D sont utilisées pour trouver le mode de liaison et les surfaces de contacts. Une fois toutes ces données récoltées, il faut un algorithme combinatoire spécial. Le plus connu est DOCK (Shoichet and Kuntz, 1991). Ces méthodes ne sont pas toujours très fiables car elles ne prennent pas en compte le fait que les structures changent quelquefois de conformation au cours du processus de liaison. Ce problème peut être diminué par des méthodes d'optimisation globale des conformations spatiales appropriées ou par des méthodes de dynamique moléculaire telles que l'approche Monte Carlo développée par Totrov *et al.*(Totrov and Abagyan, 1994). Une étude basée sur les mutations corrélées dans les alignements multiples montre qu'il est possible de détecter des relations entre protéines. La méthode du double-hybride *in silico* va même au-delà : elle vise la prédiction des régions protéiques qui vont interagir (Olmea and Valencia, 1997).

4.5.2.Un ligand comme partenaire

Au cours des 25 dernières années, une large panoplie de méthodes se sont développées pour le crible de banques de ligands et pour l'analyse des interactions moléculaires protéine-ligand. Dans le domaine du « docking » protéine-ligand, le défi est de faire face à la grande flexibilité des ligands et de modéliser les faibles interactions entre le ligand et le récepteur de la façon la plus confiante possible. Les méthodes spécifiques à ce problème sont, par exemple, celle développée par Mizutani (Mizutani *et al.*, 1994) ou encore celle de Leach (Leach, 1994). Cette dernière prend aussi en compte la flexibilité du récepteur.

4.5.3.Une molécule d'ADN ou d'ARN comme partenaire

Les recherches sur les approches bioinformatiques de « docking » protéine-ADN/ARN sont moins nombreuses. Les caractéristiques stériques déterminant les interactions protéine-ADN sont plus subtiles que celles des interactions protéine-protéine. Knegtel *et al.* ont développé MONTY, une approche Monte Carlo pour la prédiction des interactions protéine-ADN, qui intègre la flexibilité moléculaire des deux partenaires (Knegtel *et al.*, 1994).

5. L'analyse d'une séquence protéique : synthèse

Le bioinformaticien en possession d'une séquence protéique a plusieurs choix qui se présentent à lui selon le type d'informations qu'il veut acquérir sur cette séquence. Il peut récolter, *via* de nombreux programmes, des informations générales sur la séquence, telles que sa composition en nombre et en masse en acides aminés, la distribution des charges, sa localisation subcellulaire,... Une recherche de séquences homologues et leur alignement fournira déjà des informations très utiles telles que la détermination des résidus jouant un rôle prépondérant pour la structure et/ou la fonction. Il peut aussi obtenir des données structurales concernant sa séquence d'intérêt. Soit il se contente d'un modèle topologique grâce à des méthodes de prédictions de structures secondaires, soit il visera l'obtention d'un modèle tridimensionnel. Dans ce dernier cas, la première étape qu'il devra effectuer sera la recherche d'au moins un homologue dont la structure 3D est connue. S'il en trouve, il passera à la modélisation par homologie. Sinon il tentera de trouver, par « threading » ou « pseudo-threading », une protéine sur laquelle s'enfilerait bien sa séquence. Un alignement séquence-structure fiable permettra de trouver le modèle désiré. Si le « threading » ou le « pseudo-threading » ne donnent aucun « fold » correct, il reste la solution de la prédiction *ab initio,* à condition que la séquence ne soit pas trop longue. Une fois en possession du modèle, des affinages de celui-ci peuvent être réalisés par mécanique ou dynamique moléculaire. Les structures prédites peuvent éventuellement être utilisées à des fins d'analyses complémentaires. Une analyse de « docking », par exemple, entre deux partenaires supposés, requiert leur structure tridimensionnelle.

>BabR
MSAMKWETFYDAMQSADSADQLFEIVKNYAHALGFEYVSYVMSIPSLNGSLKWVPFGAFPDG
WEQRYLAQNYAEIDPLLRRGVNSIDPLIWSQNFFASAPQIWADAVKYGLKVGISQPCWAAQG
VFGLLSFVRSGPALTPGEISMLRRQLQMVTNLLHLSMYERVDVPAISCIGDVSLTLREREIL
RWTSEGKTAEIIGTILNISTRTVNFHINNVLTKLVAVNKVQAVAKARTFGLL

Figure 21 : La séquence de BabR en format FASTA avec les 3 résidus initiateurs potentiels encadrés

LES OBJECTIFS

Un système de Quorum Sensing a été identifié chez le pathogène intracellulaire *Brucella abortus*, l'organisme d'intérêt de notre laboratoire[1]. Ces bactéries du genre *Brucella* sont de petits coccobacilles de 0.5 à 0.7 µm de large et de 0.6 à 1.5 µm de long, appartenant à la classe des α_2-*Protéobactériaceae*. Elles sont asporulées, *Gram-négatives* et aérobies strictes. Le genre *Brucella* est constitué de six espèces: *B. abortus, B. melitensis, B. suis, B. neotomea, B. canis* et *B. ovis* ; chacune ayant son ou ses hôtes préférentiels. Elles sont responsables de la brucellose, une maladie de répartition mondiale se retrouvant aussi bien chez l'homme (fièvre de Malte) que chez les mammifères. Un homologue du régulateur transcriptionnel LuxR chez *B. abortus* a récemment été séquencé. Ce dernier a été obtenu de la façon suivante[2] : une banque génomique de *B.abortus*, insérée dans des plasmides multi-copies, a été transformée dans une souche-senseur d'*E.coli* c'est-à-dire une souche possédant un plasmide avec :

- un rapporteur (luminescence, couleur,...)
- un régulateur de type Quorum Sensing
- le promoteur préférentiel du régulateur placé devant le rapporteur.

Le but, au départ, était d'identifier un gène capable d'induire le rapporteur et donc d'activer le régulateur ; ce gène attendu était celui de la phéromone synthase. En effet, seule manquait la phéromone qui puisse activer le régulateur sur le plasmide. Ce qui a été obtenu est ce qu'on peut appeler un « faux positif ». Il n'avait pas d'homologie avec les protéines de type LuxI. On a pu déterminer que ce « faux positif » était en fait le gène d'un régulateur qui, probablement parce qu'il était surexprimé, activait son promoteur sans phéromone. Ce régulateur a été nommé BabR. Trois codons initiateurs ont été identifiés sur le fragment. Nous considérerons la séquence la plus longue, c'est-à-dire celle qui reprend les trois codons initiateurs (ATG) (figure 21). Cette séquence représentant BabR a une longueur de 238 acides aminés. Indépendamment de cela, avant de trouver BabR, une phéromone a été identifiée chez *B.abortus*, il s'agit d'une dDHL (cfr tableau 2), donc une HSL à chaîne acyl de 12 C et non-substituée en position 3.

Pour mieux cerner le régulateur transcriptionnel BabR de *Brucella abortus* nous nous sommes fixés comme objectif une **caractérisation structurale** de cette protéine. Nous aborderons ce travail par une approche théorique. Elle consistera en l'utilisation d'outils bioinformatiques afin d'obtenir des informations structurales de BabR. Nous tenterons d'arriver au but en suivant ces différentes étapes :

[1] URBM , Facultés Notre-Dame de la Paix de Namur
[2] Bernard Taminiau, thèse en cours

2. Premièrement, nous déterminerons la **topologie** de BabR et de ses homologues dans le but de récolter des données sur cette famille de protéines. Elles nous permettront aussi de valider les résultats obtenus ultérieurement.

3. Nous nous pencherons ensuite sur la **modélisation** du domaine C-terminal de BabR. Grâce à celui-ci, nous pourrons notamment préciser la localisation du domaine HTH. Nous entreprendrons alors une analyse des interactions HTH-ADN de différents complexes pour en retirer des informations extrapolables au modèle du domaine C-terminal de BabR. Un modèle du domaine C-terminal de LuxR ainsi que tous les autres types de données dont nous disposons à son sujet pourront également fournir de précieuses informations.

4. Enfin, nous essayerons d'obtenir des données structurales sur le domaine N-terminal et si possible, un modèle. Ce type de domaine n'a encore jamais été caractérisé structuralement auparavant. Les difficultés rencontrées seront donc plus nombreuses que pour la modélisation du domaine C-terminal.

	Banque de séquences	Séquence d'intérêt
Blastp	Protéiques	Protéique
Blastn	Nucléotidiques	Nucléotidique
Blastx	Nucléotidiques (traduites dans les six phases)	Protéique
Tblastn	Protéiques	Nucléotidique (traduite dans les six phases)
Tblastx	Nucléotidiques (traduites dans les six phases)	Nucléotidique (traduite dans les six phases)

Tableau 7 : Plusieurs programmes permettant à BLAST de faire différents types de recherche

MATERIEL ET METHODES

1. Le matériel

Les calculs de threading, la construction des modèles, les calculs de mécanique moléculaire, la visualisation et la manipulation des structures protéiques ont été effectués sur des stations de travail Silicon Graphics, modèles Octane duo et INDIGO2, fonctionnant sous les systèmes d'exploitation IRIX 6.5 et IRIX 5.3 respectivement. Les autres opérations ont été exécutées sur différents serveurs accessibles *via* le réseau Internet.

2. Les programmes

2.1. ALIGN

Il s'agit d'un programme d'alignement pairé global, c'est-à-dire qu'il va réaliser un alignement de deux séquences données qui soit optimal du début à la fin de ces séquences. L'algorithme utilisé est celui développé par E. Myers et W. Miller (Myers and Miller, 1988). Les deux paramètres modifiables ne concernent que la présentation des résultats ; ils sont ainsi pris par défaut.

http://vega.crbm.cnrs-mop.fr/bin/align-guess.cgi

2.2. BLAST

BLAST (« Basic Local Alignment Search Tool ») a pour principal objectif la comparaison d'une séquence d'intérêt à l'ensemble d'une banque de séquences (Altschul *et al.*, 1990). Elle peut être considérée comme une extension d'un alignement pairé. C'est une méthode de recherche de similarité locale c'est-à-dire qu'elle va se focaliser sur la recherche de courtes régions identiques qui serviront à réaliser l'alignement final. L'algorithme calcule ainsi toutes les paires de segments possibles de longueur fixe entre la séquence d'intérêt et les séquences de la banque. Seules les paires dont le score[1] dépasse un certain seuil sont conservées ;

[1] Celui-ci étant calculé à partir d'une matrice de scores

```
BLASTP 1.4.11 [24-Nov-97] [Build 24-Nov-97]

Reference: Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David
J. Lipman (1990). Basic local alignment search tool. J. Mol. Biol. 215:403-10.
Query= gi|631066|pir||JC2331 adrenergic receptor alpha 1A - human (572 letters)
Database: Non-redundant SwissProt (74,037 sequences; 26,661,674 total letters)
Searching.....................................................done
                                                                    Smallest
                                                                      Sum
                                                             High  Probability
Sequences producing High-scoring Segment Pairs:              Score  P(N)       N

sp|P25100|A1AD_HUMAN ALPHA-1D ADRENERGIC RECEPTOR (ALPHA ... 1513  5.5e-266  4
sp|O02666|A1AD_RABIT ALPHA-1D ADRENERGIC RECEPTOR (ALPHA ... 1465  3.9e-242  4
sp|P23944|A1AD_RAT   ALPHA-1D ADRENERGIC RECEPTOR (ALPHA ... 1416  2.0e-228  5
sp|P97714|A1AD_MOUSE ALPHA-1D ADRENERGIC RECEPTOR (ALPHA ... 1411  5.1e-220  3
sp|P15823|A1AB_RAT   ALPHA-1B ADRENERGIC RECEPTOR (ALPHA ...  650  9.2e-130  2
sp|P18841|A1AB_MESAU ALPHA-1B ADRENERGIC RECEPTOR (ALPHA ...  650  9.2e-130  2
sp|P35368|A1AB_HUMAN ALPHA-1B ADRENERGIC RECEPTOR (ALPHA ...  643  8.8e-129  2
sp|P97717|A1AB_MOUSE ALPHA-1B ADRENERGIC RECEPTOR (ALPHA ...  629  8.2e-127  2
sp|P35348|A1AA_HUMAN ALPHA-1A ADRENERGIC RECEPTOR (ALPHA ...  589  4.2e-118  2
sp|O02824|A1AA_RABIT ALPHA-1A ADRENERGIC RECEPTOR (ALPHA ...  591  1.1e-117  2

-----

sp|P25100|A1AD_HUMAN ALPHA-1D ADRENERGIC RECEPTOR (ALPHA 1D-ADRENOCEPTOR) Length = 572

 Score = 89 (41.7 bits), Expect = 5.5e-266, Sum P(4) = 5.5e-266
 Identities = 17/17 (100%), Positives = 17/17 (100%)

Query:     1 MTFRDLLSVSFEGPRPD 17
             MTFRDLLSVSFEGPRPD
Sbjct:     1 MTFRDLLSVSFEGPRPD 17

 Score = 1513 (708.4 bits), Expect = 5.5e-266, Sum P(4) = 5.5e-266
 Identities = 299/348 (85%), Positives = 299/348 (85%)

Query:    63 EDNRXXXXXXXXXXXXXDVNGTAAVGGLVVSAQGVGVGVFLAAFILMAVAGNLLVILSVA 122
             EDNR             DVNGTAAVGGLVVSAQGVGVGVFLAAFILMAVAGNLLVILSVA
Sbjct:    63 EDNRSSAGEPGSAGAGGDVNGTAAVGGLVVSAQGVGVGVFLAAFILMAVAGNLLVILSVA 122

Query:   123 CNRHLQTVTNYFIVNLAVADLLLSATVLPFSATMEVLGFWAFGRAFCDVWAAVDVLCCTA 182
             CNRHLQTVTNYFIVNLAVADLLLSATVLPFSATMEVLGFWAFGRAFCDVWAAVDVLCCTA
Sbjct:   123 CNRHLQTVTNYFIVNLAVADLLLSATVLPFSATMEVLGFWAFGRAFCDVWAAVDVLCCTA 182

Query:   183 SILSLCTISVDRYVGVRHSLKYPAIMTERKXXXXXXXXXXXXXXXXXXXXGWKEPVPPD 242
             SILSLCTISVDRYVGVRHSLKYPAIMTERK                    GWKEPVPPD
Sbjct:   183 SILSLCTISVDRYVGVRHSLKYPAIMTERKAAAILALLWVVALVVSVGPLLGWKEPVPPD 242

Query:   243 ERFCGITEEAGYAVFSSVCSFYLPMXXXXXXXXXXXXXXXSTTRSLEAGVKRERGKASEV 302
             ERFCGITEEAGYAVFSSVCSFYLPM               STTRSLEAGVKRERGKASEV
Sbjct:   243 ERFCGITEEAGYAVFSSVCSFYLPMAVIVVMYCRVYVVARSTTRSLEAGVKRERGKASEV 302

Query:   303 VLRIHCRGAATGADGAHGMRSAKGHTFRSSLSVRLLKFSREKKAAKTLAIVVGVFVLCWF 362
             VLRIHCRGAATGADGAHGMRSAKGHTFRSSLSVRLLKFSREKKAAKTLAIVVGVFVLCWF
Sbjct:   303 VLRIHCRGAATGADGAHGMRSAKGHTFRSSLSVRLLKFSREKKAAKTLAIVVGVFVLCWF 362

Query:   363 PFFFVLPLGSLFPQLKPSEGVFKVIFWLGYFNSCVNPLIYPCSSREFK 410
             PFFFVLPLGSLFPQLKPSEGVFKVIFWLGYFNSCVNPLIYPCSSREFK
Sbjct:   363 PFFFVLPLGSLFPQLKPSEGVFKVIFWLGYFNSCVNPLIYPCSSREFK 410

 Score = 101 (47.3 bits), Expect = 5.5e-266, Sum P(4) = 5.5e-266
 Identities = 17/17 (100%), Positives = 17/17 (100%)

Query:   433 VYGHHWRASTSGLRQDC 449
             VYGHHWRASTSGLRQDC
Sbjct:   433 VYGHHWRASTSGLRQDC 449

 Score = 387 (181.2 bits), Expect = 5.5e-266, Sum P(4) = 5.5e-266
 Identities = 78/93 (83%), Positives = 78/93 (83%)

Query:   480 MQAPVASRRKPPSAFREWRLLGPFRRPTTQLRAKVSSLSHKIPXXXXXXXXXXXXXSEV 539
             MQAPVASRRKPPSAFREWRLLGPFRRPTTQLRAKVSSLSHKI               SEV
Sbjct:   480 MQAPVASRRKPPSAFREWRLLGPFRRPTTQLRAKVSSLSHKIRAGGAQRAEAACAQRSEV 539

Query:   540 EAVSLGVPHEVAEGATCQAYELADYSNLRETDI 572
             EAVSLGVPHEVAEGATCQAYELADYSNLRETDI
Sbjct:   540 EAVSLGVPHEVAEGATCQAYELADYSNLRETDI 572

-----
```

Figure 22 : Résultat typique obtenu par BLAST. La première partie donne les « hits » avec leur score et N est le nombre de HSPs trouvés sur le « hit ». La deuxième partie fournit les alignements.

on les nommes HSPs (« High score Segment Pairs »). Ces « hits » sont ensuite étendus au maximum de chaque côté juqu'à ce que certains paramètres arrivent en dessous d'un certain seuil. Il en résulte des MSPs (« Maximum score Segment Pairs») qui forment la base des alignements sans « gaps » qui caractérisent les résultats de BLAST. Une amélioration de ce programme appelée Gapped BLAST (ou BLAST2) permet de relier des paires d'HSPs en un alignement plus long par des régions de moindre confiance formées de « gaps » (Altschul *et al.*, 1997).

La recherche peut se faire dans plusieurs banques et la séquence d'intérêt peut être aussi bien nucléique que protéique. En effet, plusieurs programmes que l'on peut choisir sur la page principale du site ont été développés pour ces différentes utilisations (tableau 7). Les paramètres modifiables tels que le type de matrice de scores, le nombre maximal de « hits » affichés,...seront utilisés par défaut. Les différents « hits » présentés dans les résultats sont classés par ordre décroissant de leur score. Celui-ci est la normalisation de la somme des scores donnés, à partir d'une matrice de scores, aux paires de résidus alignés. L'autre paramètre important qui caractérise chaque « hit » est la **E-value** («Expected value »). Elle représente la fréquence à laquelle on obtiendrait, par hasard, un « hit » de même score compte tenu de la longueur de la séquence d'intérêt et de la taille de la banque dans laquelle on fait la recherche. Plus la E-value est proche de zéro moins il y a de chance d'obtenir ce score par hasard et donc plus l'homologie est réelle. La E-value est exprimée de cette façon :

$E=Y \times Z \times K \times e^{-\lambda S}$

Y : longueur de la séquence
Z : taille de la banque
K et λ : paramètres de Karlin et Altschul
S : score du « hit »

Le type de résultats fournis par BLAST est illustré à la figure 22.

http://www.ncbi.nlm.nih.gov/blast/blast.cgi

2.3. PSI-BLAST

PSI-BLAST est une autre amélioration de BLAST (Altschul *et al.*, 1997). Après une recherche classique en banque de données il construit, à partir de l'alignement multiple issu des alignements pairés, un profil. Ce profil, qui est une matrice, est utilisé pour une seconde recherche. Un nouveau profil est alors construit pour les nouvelles séquences trouvées. Donc, à chaque itération, la matrice est modifiée. Ces différentes itérations successives permettent de détecter des similarités faibles mais conservant un sens biologique. Cependant, la méthode n'est pas exempte de défauts. En effet, qui dit amélioration de la sensibilité dit aussi augmentation du nombre de faux-positifs.

Les paramètres modifiables, identiques à ceux de BLAST, seront utilisés par défaut. La séquence à introduire est une séquence protéique et les résultats sont renvoyés sous la même forme que ceux de BLAST avec les mêmes paramètres (E-value, score).

http://www.ncbi.nlm.nih.gov/blast/psiblast.cgi

2.4. ClustalW

ClustalW est un logiciel d'alignement multiple progressif de séquences protéiques ou nucléiques c'est-à-dire que, contrairement aux méthodes d'alignement multiple simultané, il n'aligne pas toutes les séquences à la fois (Thompson *et al.*, 1994). Il exploite le fait que des séquences similaires sont souvent proches d'un point de vue évolutif. Il commence par regrouper les différentes séquences dans un arbre phylogénétique en les comparant deux à deux au niveau de leur similarité. Les deux séquences les plus proches sont les premières à être alignées et les autres sont ensuite alignées en suivant la hiérarchie de l'arbre phylogénétique. ClustalW utilise le positionnement des « gaps » dans les séquences proches pour guider leur insertion dans les séquences plus éloignées. L'insertion d'un « gap » entraîne une pénalité dans le score global. Cette pénalité de « gap » est d'application dans la plupart des autres méthodes d'alignement. En effet, pour pouvoir obtenir des résultats conservant un sens biologique il faut limiter l'introduction de « gaps ».

Il s'agit d'une méthode d'alignement globale : tous les résidus sont alignés quel que soit le niveau de la similarité.

Toute une série de paramètres tels que la valeur de la pénalité de « gap », la taille de la fenêtre de comparaison des séquences,... peuvent être modifiés. Dans notre cas, ils seront utilisés par défaut.

http://dot.imgen.bcm.tmc.edu:9331/multi-align/Options/clustalw.html

2.5. Match-Box

Match-Box est une méthode d'alignement multiple simultané dont le but est de rechercher des segments similaires dans un ensemble de protéines et de les faire correspondre dans un alignement (Depiereux and Feytmans, 1992), (Depiereux *et al.*, 1997). Match-Box est composé de deux sous-programmes : l'un est appelé EXPLORE et l'autre ALIGN[1]. L'algorithme d'EXPLORE, effectué dans un premier temps, détermine si les similarités observées se démarquent du hasard. ALIGN est l'alignement en soi. L'opération d'alignement qui suit se décompose

[1] Il n'a rien à voir avec le programme d'alignement pairé du point 2.1.

```
>BabR
--------------------------------------------------
--------------------------------------------------
--------------------------------lqmvtnllhlsmyervdv
paisci----------------------------------------------
--------------------------------------------------
--------------------------gdvsltlrereilrwtsegktae
iigt----------------------------ilnistrtvnfHINNVL
TKLVAVNKVQAVAKARTFGLL----------------
>CepR
--------------------------------------------------
-----------------------------------MELRWQDAYQQFS
AAEDEQQLFQRIAAYSKRLGFEYCCYGIRVPLpvskpavaifdtypdgwm
ahyqaqNYIEIDSTVRDGALNTNMIVWPDVDRIDPCPLWQDARDFGLSVG
VAQSSWAARGAFGLLSIARHADRLTPAEINMLTLQTNWLANLSHSLMSRF
MVPKLSPA-------------------agvtltardrevlcwtaegktac
eigq----------------------------ilsisertvnfHVNNIL
EKLGATNKVQAVVKAISAGLIEAP--------------
>SolR
--------------------------------------------------
-----------------------------------MEPDFQDAYHAFR
TAEDEHQLFREIAAIARQLGFDYCCYGARMPLpvskpavaifdtypagwm
qhyqasGFLDIDPTVRAGASSSDLIVWPVSIRDDAARLWSDARDAGLNIG
VARSSWTAHGAFGLLTLARHADPLTAAELGQLSIATHWLANLAHTLMSPF
LVPQLVPE-------------------snavlttrerevlcwtgegktay
eigq----------------------------ilrisertvnfHVNNVL
LKLAATNKVQAVVKAIATGLI----------------
>PhzR
--------------------------------------------------
-----------------------------------MHDEREGYLEILS
RITTEEEFFSLVLEICGNYGFEFFSFGARAPFpltapkyhflsnypgewk
sryiseDYTSIDPIVRHGLLEYTPLIWNGEDFQENRFFWEEALHHGIRHG
WSIPVRGKYGLISMLSLVRSSESIAATEILEKESFLLWITSMLQATFGDL
LAPRIVPE-------------------snvrltaretemlkwtavgktyg
eigl----------------------------ilsidqrtvkfHIVNAM
RKLNSSNKAEATMKAYAIGLLN----------------
>PHZR_PSEF
--------------------------------------------------
----------------------------MFKMELGQLLGWDAYFYSIFA
QAMNMEEFTVVALRALRELRFDFFAYGMCSVTpfmrpktymygnypehwl
qryqaaNYALIDPTVKHSKVSSAPILWSNELFRNCPDLWSEANDSSLCHG
LAQPSFNTQGRVGVLSLARKDNAISLQEFEALKPVTKAFAAAALEKISAL
ETDVRAFN-------------------tdvefserecdvlrwtadgktse
eigv----------------------------imgvctdtvnyHHRNIQ
RKIGASNRVQAVSYAVALGYI----------------
>PhzR
--------------------------------------------------
-------------------------------MELGQQLGWDSYFYNIFA
RTMDMQEFTAVTLRVLRELRFDFFAYGMCSVTpfmrprtcmygnypedwv
qryqaaNYAVIDPTVKHSKVSSAPILWSNELFRGCPDLWSEANDSNLCHG
LAQPSFNAQGRVGMLSLARKDNPISLQEFEALKLMTKAFAAAIHEKISEL
ESDVRVFN-------------------tdvefsgrecdvlrwtadgktse
eigv----------------------------imgvctdtvnyHHRNIQ
RKIGASNRVQAVSYAVAMGYI----------------
```

Figure 23 :Séquences sous format FASTA alignées selon Match-Box. Chaque trait représente un « gap »

elle aussi en plusieurs étapes successives : le *scanning* (paramétrisation du système), le *matching* (définition des boîtes de segments similaires susceptibles de se retrouver dans l'alignement final) et le *screening* (triage des boîtes afin de réaliser l'alignement final).

Certains paramètres peuvent être modifiés par l'utilisateur, comme par exemple le type de matrice de scores, mais ils seront utilisés par défaut dans notre cas. Les résultats nous parviennent sous forme de quatre fichiers : un fichier de confirmation du lancement du travail, un fichier EXPLORE, un fichier ALIGN et un fichier reprenant les différentes séquences alignées sous format FASTA (figure 23). Un aspect intéressant de Match-Box est l'attribution d'un score de confiance à chaque position alignée. Ceci permet d'évaluer la fiabilité de l'alignement proposé. Le meilleur score attribué est 1 et un score supérieur à 5 indique un alignement qui ne se démarque pas du hasard. Il s'agit d'un alignement local : en dehors des zones reconnues comme similaires, les résidus ne sont pas alignés.

http://www.fundp.ac.be/sciences/biologie/bms/matchbox_submit.html

2.6. PHD

PHD est un logiciel constitué de plusieurs programmes réunis en un « super-outil ». On y trouve des programmes permettant de prédire les structures secondaires (PHDsec), l'accessibilité au solvant (PHDacc), les hélices transmembranaires (PHDhtm), les régions formant des « coiled coils » (COILS), les motifs fonctionnels annotés par des experts (PROSITE), les domaines (ProDom), les cystéines liées (CYSPRED) et la tendance qu'a la protéine d'intérêt à être globulaire (GLOBE) (Rost and Sander, 1993). PHDsec se base sur un alignement multiple pour améliorer la confiance de ses prédictions d'hélices, de brins et de boucles. Cet alignement est réalisé soit par PHDsec, soit il est fourni par l'utilisateur. Le principe utilisé par ce programme est celui des réseaux neuronaux. PHDsec utilise aussi les informations données par les méthodes décrites ci-dessus pour améliorer la prédiction des structures secondaires. La performance de PHD atteint 71,4% d'efficacité c'est-à-dire, qu'en moyenne, 71,4% des résidus sont assignés à la bonne structure secondaire. Elle est la plus populaire des méthodes de prédictions de structures secondaires bien que d'autres méthodes telles que celles décrites ci-dessous commencent sérieusement à la concurrencer. Un programme de prédiction de repliement appelé TOPITS est accessible sur ce même site (voir point 2.14).

Les nombreux paramètres modifiables des différents programmes seront gardés par défaut. Les résultats de l'ensemble des programmes que l'on a sélectionnés sur la page principale sont renvoyés dans un seul fichier. En ce qui concerne les prédictions de structures secondaires, elles sont accompagnées de coefficients de certitude allant de 0 à 9, ce dernier étant la certitude la plus élevée.

```
protein:        predict         length      238


                 ....,....1....,....2....,....3....,....4....,....5....,....6
        AA       |MSAMKWETFYDAMQSADSADQLFEIVKNYAHALGFEYVSYVMSIPSLNGSLKWVPFGAFP|
        PHD sec  |      HHHHHHHHHH  HHHHHHHHHHHHHH   EEEEEEEE          EEEE    |
        Rel sec  |9776415898988862113899999999988336325698742677899995578727 91|
detail:
        prH sec  |0111247888888875446899999999988631211100000000000000000000004|
        prE sec  |0001100000000000000000000000000000025678876521110000268873100|
        prL sec  |9877642001001123553100000000011367532100124787889996210158 95|
subset: SUB sec  |LLLL..HHHHHHHHH....HHHHHHHHHHHH..L..EEEEE..LLLLLLLLLEEEE.LL.|
accessibility
3st:    P_3 acc  |eebbebeebbeebeeeeeeeebbebbeeebeebebebbbbbbbbebeeeeeeebbbbbbbb|
10st:   PHD acc  |9700606600660676777770060067607606060000000606777677000000000|
        Rel acc  |0200110041102100100212202412031150404453410011110101315500 01|
subset: SUB acc  |........b.............b......b.b.bbb.b............bb....|


                 ....,....7....,....8....,....9....,....10...,....11...,....12
        AA       |DGWEQRYLAQNYAEIDPLLRRGVNSIDPLIWSQNFFASAPQIWADAVKYGLKVGISQPCW|
        PHD sec  |HHHHHHHHHH       E              HHHHH HHHHHHHHHH    E  EE    |
        Rel sec  |8999999984544378111243269998421235662138999997215732112111 12|
detail:
        prH sec  |8999999986221100112210010000001367775468999988552122211111 10|
        prE sec  |0000000000002210343223310000243200000000000000000002333344443|
        prL sec  |1000000013665578444465579898654432223531000001447754345344 45|
subset: SUB sec  |HHHHHHHHH.L...LL.......LLLLL.....HHH...HHHHHHH..LL..........|
accessibility
3st:    P_3 acc  |eebbe b bebbebbbbbbeebeeeebbbbbbeebbeebeebbeebee bbbebbbbbb  |
10st:   PHD acc  |7600750507007000000660777700000076007707600760675000600000 04|
        Rel acc  |2150203101210050345112110001304101402111263216110351164213 20|
subset: SUB acc  |..b..........b..bb..........b...b......b...b....b..bb.....|
```

Figure 24 : Résultat PHD. La ligne « PHD sec » donne la prédiction pour chaque résidu et la ligne « Rel sec » représente la confiance donnée à chaque prédiction. La ligne « SUB acc » prédit l'accès au solvant

PHDacc utilise les symboles b (« buried ») et e (« exposed ») pour déterminer les résidus enfouis et exposés respectivement (figure 24).

http://www.embl-heidelberg.de/Services/sander/predictprotein/

2.7. PROF

PROF est une méthode statistique[1] de prédictions de structures secondaires (King *et al.*, 1997). Mais contrairement à PHD qui utilise des statistiques non-linéaires, ne fournissant pas le principe gouvernant leurs prédictions[2], PROF a pour but de clarifier les statistiques utilisées tout en améliorant les prédictions fournies par des méthodes statistiques plus simples. Pour en arriver là, il ne se base pas simplement sur la séquence de résidus mais aussi sur les concepts sous-jacents, comme par exemple les profils d'hydrophobicité associés à différents types de structures secondaires. De plus, l'utilisation d'alignements de séquences homologues permet de récolter des informations supplémentaires telles que l'identification de régions variables, celles-ci ne formant probablement pas d'hélices ou de brins mais plutôt des boucles plus sujettes à la variation. PROF atteint une efficacité de 70,1% bien qu'elle soit supérieure à PHD pour des protéines allant de 90 à 170 résidus.

Tous les paramètres seront utilisés par défaut. Dans le tableau de résultats, trois probabilités[1] sont attribuées aux résidus. Elles permettent d'évaluer la fiabilité des prédictions : plus le chiffre s'approche de 1 pour un type de structures secondaires et plus sûr sera le résultat.

http://www.bmm.icnet.uk/servers/prof/?p=form

2.8. PREDATOR

PREDATOR est un programme qui prédit des structures secondaires à partir d'une séquence unique ou d'un set de séquences homologues. Dans ce dernier cas, il n'utilisera pas un alignement multiple mais les alignements pairés locaux de chaque séquence du set avec la séquence d'intérêt, pour en retirer des informations supplémentaires. L'algorithme de base repose sur la reconnaissance de paires potentielles de résidus allant former des ponts hydrogène. Cette méthode implique des statistiques dérivées de banques de données sur l'occurrence des différents types de résidus dans la formation des ponts H entre brins β. Les hélices α sont aussi reconnues sur base de l'occurrence des acides aminés dans les paires reliées par des ponts H (résidu i et i+4). L'algorithme atteint une efficacité de 68% pour

[1] Les descriptions qui suivent sont celles du programme DSC, la version précédant PROF, car l'article dont la référence est indiquée sur le site Internet n'est pas encore disponible (Ouali and King,). La référence de King, ci-dessus, est celle du programme DSC.
[2] Puisque PHD utilise le principe des réseaux neuronaux.
[3] Une pour chaque type de structures secondaires : hélice, brin, boucle

une séquence seule, mais de 75% lorsque sont rajoutées les informations provenant des alignements pairés.

Tous les paramètres seront utilisés par défaut. On peut envoyer un alignement ClustalW à PREDATOR mais il le considérera comme un set non-aligné. Les résultats sont présentés sans coefficient de confiance.

http://www.embl-heidelberg.de/cgi/predator_serv.pl

2.9. JPred2

Le serveur Internet JPred2 permet la prédiction de structures secondaires (Cuff *et al.*, 1998)[1]. Il autorise l'envoi d'une séquence seule ou d'un alignement multiple et il renvoie les prédictions de sept algorithmes de prédictions de structures secondaires : NNSSP, PROF, PHD, MULPRED, PREDATOR, ZPRED et JNET. Toutes ces méthodes exploitent les informations contenues dans un alignement multiple. Une prédiction consensus de ces méthodes est également renvoyée. Dans le cas de l'envoi d'une séquence d'intérêt unique, JPred2 utilise une recherche PSI-BLAST pour obtenir les séquences homologues à la séquence d'intérêt et les alignements multiples sont réalisés par une version plus rapide de ClustalW. Les profils générés par PSI-BLAST améliorent l'efficacité de JNET ; il est donc recommandé de n'envoyer qu'une seule séquence. JPred2 prédit aussi les régions « coiled coils » à l'aide de COILS et MULTICOIL. Sans inclure la nouvelle méthode JNET, la prédiction consensus est efficace à 72,9%. Les données ne sont pas encore disponibles mais on sait déjà que l'introduction de JNET a fait augmenter ce chiffre.

Par défaut, nous exécutons uniquement JNET. En effet l'utilisation de JNET seule ne diminue pas beaucoup l'efficacité de la prédiction par rapport à celle du consensus, tout en diminuant le temps calcul. De plus, JNET est elle-même une méthode consensus « home-made » de JPred2. Dans notre cas, nous n'avons pas demandé l'exécution des autres méthodes dans le consensus final de JPred2 puisque nous les inclurons nous-même par la suite dans un consensus fait à la main. Les autres paramètres seront aussi utilisés par défaut.

http://jura.ebi.ac.uk:8888/

[1] Cette référence est celle de JPred, premier du nom, mais la différence entre les deux versions se résume à quelques programmes en plus (JNET, COILS, MULTICOIL). Pour l'instant, aucun article sur JPred2 n'est disponible.

2.10. PSIpred

Ce serveur propose trois méthodes : PSIpred, une méthode de prédiction de structures secondaires à laquelle nous nous intéressons ici ; MEMSAT2, qui prédit les topologies trans-membranaires ; et GENTHREADER, une méthode de prédiction de repliements (voir point 2.12).

PSIpred incorpore le principe des réseaux neuronaux (Jones, 1999b). Ceux-ci réalisent une analyse sur le profil généré par PSI-BLAST contrairement à ceux de PHD qui utilisent le profil généré par le programme d'alignement multiple MAXHOM. Son score de 77% d'efficacité est le meilleur score recensé pour les méthodes de prédictions de structures secondaires.

Une séquence unique est envoyée. Les quelques paramètres modifiables seront gardés par défaut. Les résultats sont présentés accompagnés d'une évaluation allant de 0 à 9 pour chaque prédiction ; 9 étant le meilleur score.

http://insulin.brunel.ac.uk/psipred/

2.11. 3DPSSM

3DPSSM est un serveur dédié à la prédiction de la structure tridimensionnelle d'une protéine d'intérêt et de sa fonction probable (« fold recognition ») (Kelley *et al.*, 2000). Elle se base sur une banque de protéines de structure connue sur chacune desquelles la séquence d'intérêt va être « enfilée ». Cette dernière opération est évaluée par un score de compatibilité. Ce score est déterminé par plusieurs composants :

-la comparaison avec le 1DPSSM qui est le profil de séquences construit à partir de l'alignement d'homologues relativement proches de la protéine de structure 3D connue. Ce profil est obtenu par PSI-BLAST

-la comparaison avec le 3DPSSM qui est un autre profil plus général contenant des homologues plus éloignés obtenus par superposition de la protéine de structure connue et les autres protéines de la même superfamille. Les structures se superposant à la protéine de structure connue avec un RMS < à 6Å sont ajoutées à l'alignement utilisé pour faire le profil.

-la superposition plus ou moins bonne d'éléments de structures secondaires. Ces derniers sont prédits à l'aide de PSIpred.

-la tendance qu'ont les résidus de notre séquence à occuper différents niveaux d'accessibilité au solvant.

Les paramètres seront gardés par défaut. Dans la présentation des résultats, les 20 structures les plus probables sont classées de la meilleure à la plus mauvaise. Un paramètre important, attribué à chaque structure trouvée, est la E-value qui

permet de donner une confiance à la prédiction. Ce type de score a déjà été évoqué précédemment dans le point 2.2.

http://www.bmm.icnet.uk/~3dpssm/

2.12. GENTHREADER

GenTHREADER est une méthode de reconnaissance de repliements ("fold") confiante et rapide (Jones, 1999a). Pour cette raison, elle est particulièrement bien adaptée à l'annotation automatique de génomes. L'algorithme utilise un profil généré par PSI-BLAST pour déterminer un alignement séquence-structure de la séquence d'intérêt avec chaque élément de la banque de séquences de structure 3D connue. Pour chaque alignement séquence-structure, les potentiels d'énergie pairés et les termes de solvatations qui sont sommés , ainsi que les scores de l'alignement, sont présentés à un réseau neuronal entraîné sur les similarités structurales trouvées dans CATH. Il détecte les combinaisons favorables entre les sommes d'énergie et les scores de l'alignement. L'utilisation d'un profil permet d'éliminer en partie les prédictions faussement positives puisque les relations évolutives sont prises en compte.

GenTHREADER est accessible sur le serveur de PSIpred. Il suffit de sélectionner GenTHREADER plutôt que PSIpred et la séquence d'intérêt peut être envoyée en gardant, dans notre cas, les quelques paramètres modifiables par défaut. Seules les 10 meilleures structures sont retenues dans les résultats. Une probabilité estimée d'avoir la structure correcte est donnée pour chacun de ces 10 repliements.

http://insulin.brunel.ac.uk/psipred/

2.13. THREADER 2.5

THREADER 2.5 est un programme de « threading » tournant sur Silicon Graphics, que l'on peut télécharger à partir du site donné ci-dessous (Jones *et al.*, 1999). La séquence d'intérêt va être enfilée d'une façon optimale, sur chaque « fold » de la banque en utilisant un algorithme de programmation dynamique. L'énergie de chaque enfilement possible est calculée en sommant les interactions pairées et les paramètres de solvatation. L'entièreté de la banque de « folds » est ensuite remaniée en classant ceux-ci par ordre d'énergie croissante, le « fold » de plus basse énergie étant la prédiction la plus probable. L'amélioration de THREADER2.5 par rapport à la version précédente est l'addition de nouvelles caractéristiques permettant aux informations provenant de la séquence ainsi que des structures secondaires, d'être considérées dans le processus de reconnaissance de « folds » (ou repliements). Les informations provenant de la séquence sont reprises dans la fonction calculant le score séquence-structure de chaque

prédiction. Un certain poids est ainsi donné à une similarité de séquence éventuelle. Au contraire, les informations tirées des prédictions de structures secondaires ne sont pas incorporées à cette fonction. Cependant, elles jouent un rôle dans l'alignement séquence-structure. L'algorithme va ignorer les alignements de plus faible énergie au profit d'alignements de plus haute énergie qui sont en concordance avec les structures secondaires ; c'est-à-dire qu'il n'alignera pas une hélice avec un brin.

Le score total pour chaque « fold » est normalisé en un Z-score :
Z>3.5 : sûrement correct
3.5>Z>2.0 : éventuellement correct
Z<2.0 : probablement pas correct

http://globin.bio.warwick.ac.uk/~jones/threader.html

2.14. TOPITS

TOPITS est accessible sur le même site que PHD. Cette méthode de « pseudo-threading » ou "fold recognition" repose sur le principe suivant : détecter des motifs similaires de structures secondaires et d'accessibilité au solvant entre une séquence d'intérêt et une séquence de structure 3D connue (Rost, 1995). Les structures secondaires de la séquence d'intérêt sont prédites par PHD, et TOPITS va aligner le profil 1D de cette séquence avec celui de tous les « folds » de la banque de structures par un algorithme de programmation dynamique. Ce profil 1D est constitué des éléments de structures secondaires et d'accessibilité au solvant.

Les résultats sont présentés avec les 20 meilleurs « folds » classés dans l'ordre décroissant de leur score. Un score est attribué à chaque alignement selon le pourcentage de profils identiques. Ce score est ensuite normalisé en un Z-score appelé ZALI dont les seuils ont été décrits pour la méthode précédente. Les paramètres modifiables sont les mêmes que ceux de PHD. Ils seront également gardés par défaut.

http://www.embl-heidelberg.de/Services/sander/predictprotein/

2.15. PSI-BLAST-BORK

Il s'agit d'une autre méthode de prédiction de structures tertiaires. PSI-BLAST-BORK qui se base sur la recherche de séquences homologues est en fait une extension de PSI-BLAST (Huynen *et al.*, 1998). En effet, la séquence d'intérêt est comparée à l'aide de PSI-BLAST à une version de la banque NR qui contient uniquement les entrées de PDB. Le seuil de la E-value est limitée à 0,001 et le nombre maximal d'itérations est fixé à 5. Les régions à « coiled coils », les régions

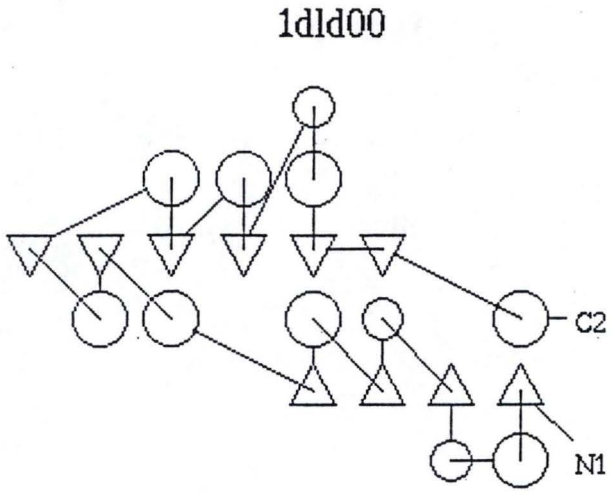


Figure 25 : Diagramme TOPS de l'entrée PDB 1dld. Deux plans β parallèles pris en sandwich entre 4 couches d'hélices α

trans-membranaires et les régions de faible complexité (LCR) sont automatiquement filtrées par les programmes COILS, Tmpred et SEG. S'ils étaient gardés, ces éléments pourraient provoquer des erreurs de prédictions (faux positifs).

Les paramètres cités ci-dessus sont les seuls à être modifiés. Les résultats sont présentés sous la même forme qu'un « output » de PSI-BLAST avec un graphique montrant les parties alignées.

http://dove.EMBL-Heidelberg.DE/3D/

2.16. TOPS

TOPS est un programme de simplification de la représentation de structures tertiaires. La structure 3D est schématisée en un diagramme à deux dimensions. Celui-ci se présente comme une succession d'éléments de structures secondaires dont la position spatiale relative et la direction sont indiquées. Les brins sont des triangles et les hélices des cercles. Des triangles dont l'orientation pointe alternativement dans un sens puis dans l'autre déterminent un plan β anti-parallèle. Lorsqu'ils pointent dans la même direction, ils déterminent un plan β parallèle (figure 25). Deux protéines adoptant la même topologie ne présentent pas nécessairement la même structure tridimensionnelle bien qu'il existe une relation structurale entre les deux. Un algorithme pour la génération automatique de topologies, à partir de structures 3D, a été développé par Flores et ses collaborateurs en 1994 ; il a récemment été amélioré (Westhead *et al.*, 1999). Une banque de topologies est aussi disponible sur le site de ce programme.

http://www3.ebi.ac.uk/tops/

2.17. SWISSMODEL

SWISSMODEL (Guex and Peitsch, 1997) est un serveur de modélisation automatisée par homologie. Il commence par une recherche d'homologues de structure 3D connue dans la banque ExPDB (banque dérivée de PDB) *via* BLAST2. La deuxième étape consiste à sélectionner les séquences « template » possédant une identité > à 25 % avec la séquence « target ». De plus, les domaines pouvant être modélisés séparément à partir de différents « template », seront détectés. C'est l'algorithme SIM qui se charge de cette étape. La suivante va générer les fichiers nécessaires à ProModII. Celui-ci réalisera, pour la quatrième étape, le modèle atomique tridimensionnel à partir d'un alignement séquence-structure. Enfin, la cinquième et dernière étape consiste en une évaluation du modèle par Gromos96.

Il est possible de fournir, avec la séquence « target », un fichier PDB. SWISSMODEL sautera alors la première étape. Ceci est intéressant si nous possédons déjà un homologue de structure 3D connue. Les paramètres modifiables seront gardés par défaut. Un fichier PDB du modèle est renvoyé par courrier électronique. ProModII calcule un facteur de confiance du modèle, le facteur B, qui est inclus dans le fichier PDB. Il est déterminé, entre-autres, par le nombre de structures « template » utilisées pour la modélisation et par la déviation du modèle par rapport à la structure.

http://www.expasy.ch/swissmod/SWISS-MODEL.html

2.18. MODELLER 4

MODELLER est un programme dont l'objectif est la modélisation de structures 3D de protéines en respectant les contraintes spatiales (Sanchez and Sali, 1997). Il est utilisé principalement pour la modélisation par homologie : l'utilisateur fournit un alignement de la séquence à modéliser avec la structure connue correspondante et MODELLER calcule automatiquement un modèle atomique. Il prend en compte les contraintes spatiales dérivées de la structure homologue connue et de son alignement avec la séquence « target ». MODELLER peut aussi réaliser des comparaisons multiples de séquences de protéines et/ou de structures, grouper et classifier un set de protéines, faire une recherche de séquences ou de structures en banques de données, optimiser le modèle,... Ces nombreuses fonctions peuvent être activées par des lignes de commandes dans un fichier « nom.top ». Celui-ci doit être accompagné de deux autres fichiers pour lancer le programme : un fichier « nom.atm » qui contient les coordonnées du « template » et un fichier « nom.ali » reprenant l'alignement séquence-structure. Six fichiers de résultats sont crées ; les plus importants sont le fichier « nom.B999... » contenant les coordonnées du modèle et le fichier « nom.log » qui décrit la succession des étapes et les erreurs éventuelles rencontrées lors du processus de modélisation. MODELLER fonctionne sur un ordinateur tournant sous UNIX. Des informations supplémentaires peuvent être trouvées sur le site ci-dessous.

http://guitar.rockefeller.edu/modeller/modeller.html

2.19. Insight II

Insight II est un ensemble de programmes de modélisation moléculaire, qui consiste en un groupe de modules accessibles *via* une interface graphique (1995b, guide de l'utilisateur). Il tourne sur une machine fonctionnant sous le système d'exploitation UNIX, une Silicon Graphics dans notre cas. Le programme permet de créer, de modifier, de manipuler, de visualiser et d'analyser les systèmes moléculaires et les données en relation. Un module important que nous avons

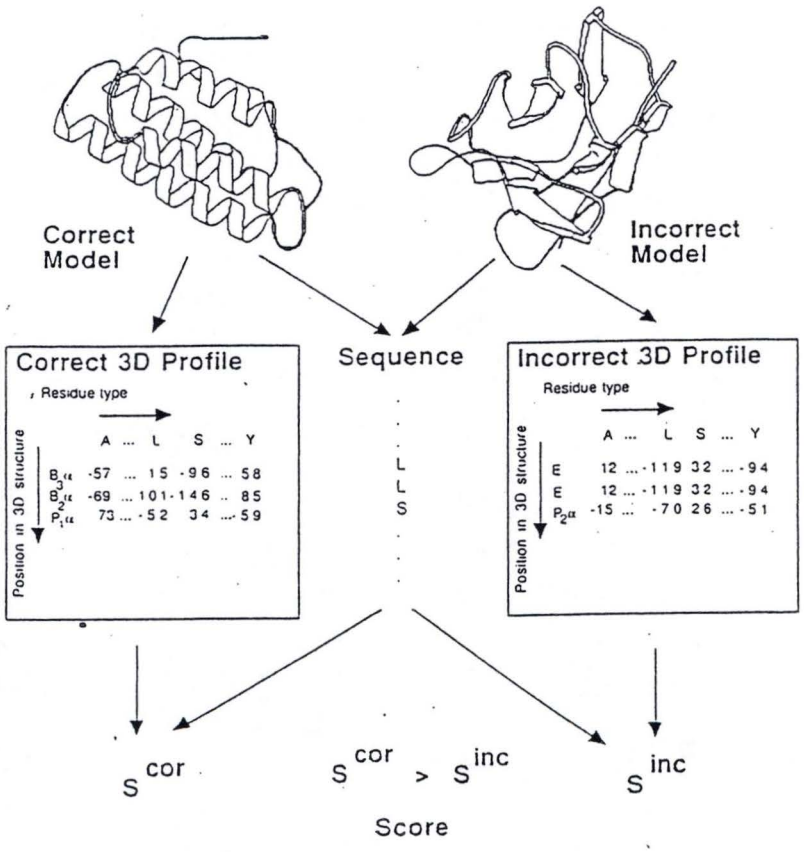


Figure 26 :Comparaison de l'attribution de scores 3D-1D à un « fold » correct et à un « fold » incorrect. Le « score S » final sera plus grand pour le « fold » correct que pour le « fold » incorrect

utilisé est DISCOVER. Celui-ci est un programme de simulation moléculaire qui utilise un champ de force pour réaliser des opérations de mécanique et de dynamique moléculaire (1995a, guide de l'utilisateur). Un champ de force est la forme mathématique de l'énergie potentielle d'un système. Discover supporte 4 familles de champ de force choisis selon le type de molécule sur laquelle on travaille. Celui que nous utiliserons est le champ de force CVFF qui a été paramétrisé pour reproduire les propriétés des peptides et des protéines. C'est le champ de force par défaut du programme. Des informations supplémentaires peuvent être trouvées sur le site ci-dessous.

http://www.csc.fi/chem/progs/insightII.html

2.20. VERIFY3D

VERIFY3D permet d'évaluer un modèle (Luthy *et al.*, 1992). La méthode mesure la compatibilité d'un modèle protéique avec la séquence, en utilisant les caractéristiques tridimensionnelles : la position de chaque résidu dans le modèle 3D est caractérisée par son environnement et est représentée par une rangée de 20 chiffres. Ces derniers correspondent aux préférences statistiques (appelées scores 3D-1D) des 20 acides aminés pour cet environnement. L'environnement de chaque résidu est défini par trois paramètres : la partie du résidu qui est enfouie, la fraction de la surface des chaînes latérales couverte par des atomes polaires et la structure secondaire locale. Le graphique généré par cette évaluation est le résultat de la somme des scores 3D-1D à l'intérieur d'une fenêtre de 21 résidus qui se déplace sur la séquence. Un score négatif indique un problème local. Le score total est dit « score S » (figure 26). Quelques exemples de profils sont montrés à la figure 27. La seule opération que nous devons effectuer est l'envoi du fichier PDB de notre modèle.

http://www.doe-mbi.ucla.edu/verify3d.html

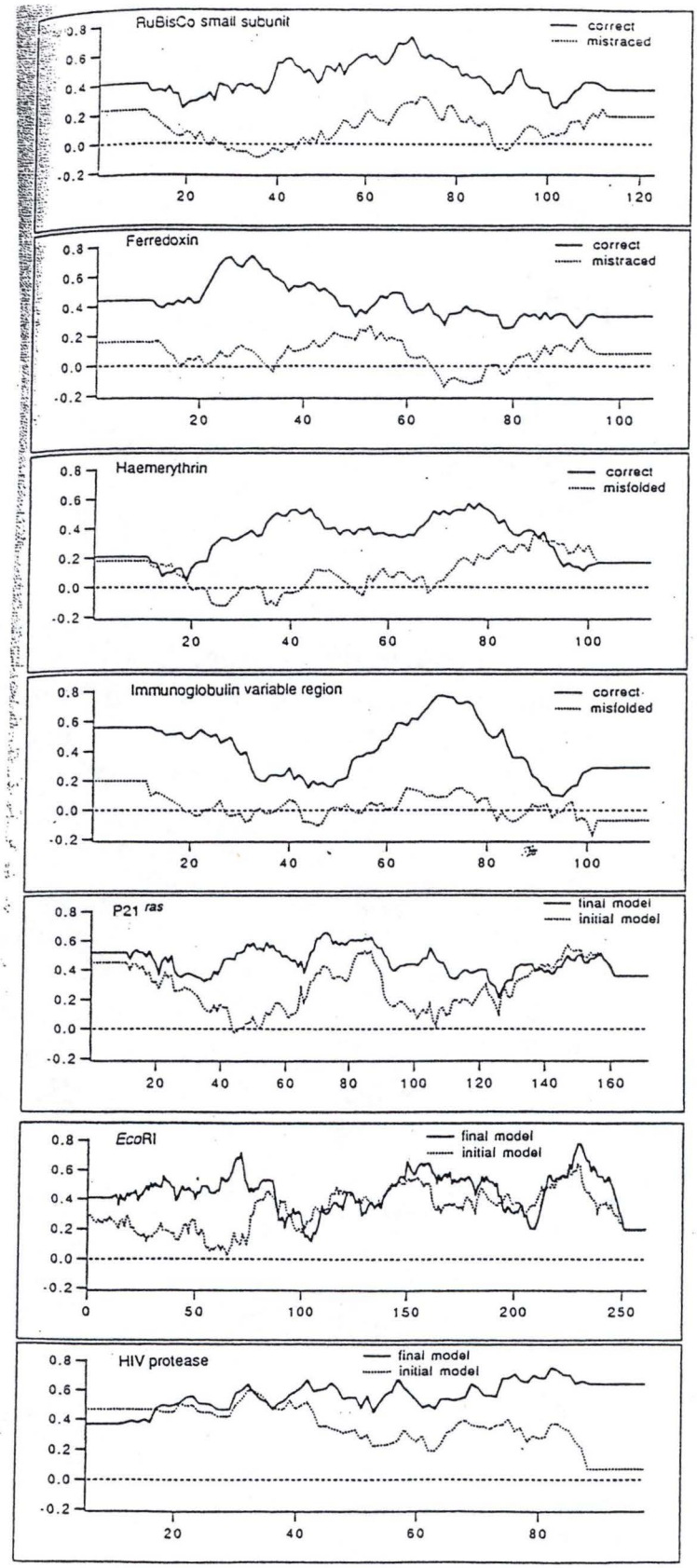


Figure 27 :Profil Verify 3D de plusieurs modèles incorrects ou partiellement incorrects ; l'axe vertical donne le score 3D-1D moyen pour les résidus dans une fenêtre de 21 résidus ; les scores des 9 premières et des 9 dernières positions n'ont aucune valeur

RESULTATS ET DISCUSSION

La stratégie générale de ce travail consiste à obtenir d'abord un petit ensemble d'homologues assez similaires à BabR dont nous déterminerons quelques caractéristiques par des méthodes d'alignement multiple et par des méthodes de prédictions de structures secondaires. LuxR, une protéine sur laquelle de nombreuses informations sont disponibles, est inclue à ce petit ensemble d'homologues. Un consensus sera tiré des méthodes de prédictions de structures secondaires. Nous allons donc obtenir des informations générales sur les homologues les plus proches de BabR. Pour la suite des opérations, BabR sera scindé en deux domaines. Une modélisation par homologie sera effectuée préférentiellement sinon nous nous tournerons vers d'autres méthodes de prédictions de structures tertiaires. L'examen plus détaillé du domaine HTH de BabR sera réalisé par une analyse de différents complexes HTH-ADN de structure 3D résolue.

1. Recherche d'homologues pour BabR

La première opération effectuée est la recherche de séquences homologues à notre protéine d'intérêt BabR de *Brucella abortus*. Rappelons qu'il s'agit d'un régulateur transcriptionnel activé par une phéromone, intervenant dans un phénomène dépendant de la densité cellulaire appelé Quorum Sensing. Rappelons aussi qu'une phéromone pouvant intervenir dans le Quorum Sensing été identifiée chez *B.abortus*. Il s'agit d'une dDHL mais nous ne sommes pas certains qu'elle se lie à BabR. Le régulateur de Quorum Sensing le plus étudié et donc celui sur lequel nous possédons le plus de données est LuxR de *Vibrio fischeri*.

Une recherche *via* BLASTP dans la banque de séquences non redondantes NR a abouti à la découverte de 109 homologues. Ce sont tous des régulateurs de transcription et la plupart d'entre eux sont des régulateurs appartenant à la superfamille LuxR (ou impliqués dans le Quorum Sensing). Dans le but de réaliser un premier triage nous avons sélectionné les séquences ayant un score inférieur à 0,01. Pour détecter les redondances et les séquences trop similaires nous avons construit une matrice contenant le pourcentage d'identité entre toutes ces séquences. La matrice est créée par une extension d'ALIGN qui aligne toutes les séquences deux à deux et détermine leur identité. Les séquences qui partagent une identité supérieure à 95% avec une autre ont été écartées. Ainsi, les séquences se retrouvant en plusieurs exemplaires avec deux ou trois résidus mutés sont éliminées mais les séquences similaires d'organismes proches sont gardées. Ceci permet de diminuer le temps

calcul des méthodes employées plus tard et d'éviter les biais dans les résultats. On obtient un ensemble de 33 homologues appartenant à la famille LuxR, LuxR se retrouvant bien parmi ceux-ci.

Nous nous intéresserons plus loin à certaines caractéristiques structurales des homologues les plus proches de BabR ; c'est pourquoi, nous avons encore subdivisé le set en trois groupes de séquences:
- le groupe des 9 homologues dont l'identité avec BabR est supérieure à 30%
- le groupe des 9 homologues dont l'identité avec BabR est supérieure à 30% + LuxR
- le set en entier (33 séquences)

L'intérêt du second groupe par rapport au premier est le fait qu'il contient LuxR (25% d'identité avec BabR d'après l'alignement donné par ALIGN) et donc les informations dont nous disposons à son sujet pourront éventuellement être extrapolées aux autres séquences du groupe à partir d'un alignement multiple de ces séquences. Ainsi, l'idéal serait de pouvoir utiliser le second groupe dans les étapes suivantes. Seulement, l'ajout de LuxR qui n'est pas un homologue proche de BabR (<30% d'identité), pourrait perturber l'alignement par rapport à celui donné pour les séquences du premier groupe. Nous vérifierons cela dans le point 2. Nous comparerons aussi les alignements des trois groupes pour déterminer les régions les plus conservées.

Une autre classification pourrait donner d'autres informations. C'est pourquoi, nous avons regroupé les 33 homologues selon le type d'HSL qu'ils lient. D'après les données bibliographiques dont nous disposions, nous avons pu classer 19 homologues :

type d'HSL	**Les régulateurs correspondants**
C8HSL	CepR, SolR, YtbR
C6HSL	PhzR psear, YenR, PhzR pseae, YpsR
C4HSL	RhlR pseae, AhyR
7cis-C14HSL	CerR
3-oxo-C6HSL	LuxR, EsaR, ExpR, CarR, YpsR
3-oxo-C10HSL	VanR
3-oxo-C12HSL	LasR
3-oxo-C8HSL	TraR
3-OH-7cis-C14HSL	RhiR

Nous observons que CepR et SolR, les deux homologues les plus similaires à BabR, lient le même type de phéromone. La phéromone identifiée chez *B.abortus* ressemble à celle qui lie CepR et SolR puisque toutes les deux sont non-substituées et leur chaîne acyl a presque la même longueur (8 C pour celle se liant à CepR et SolR et 12 C pour celle trouvée chez *B.abortus*). Ceci est un argument pour dire que la dDHL trouvée chez *B.abortus* est bien la phéromone qui se lie à BabR.

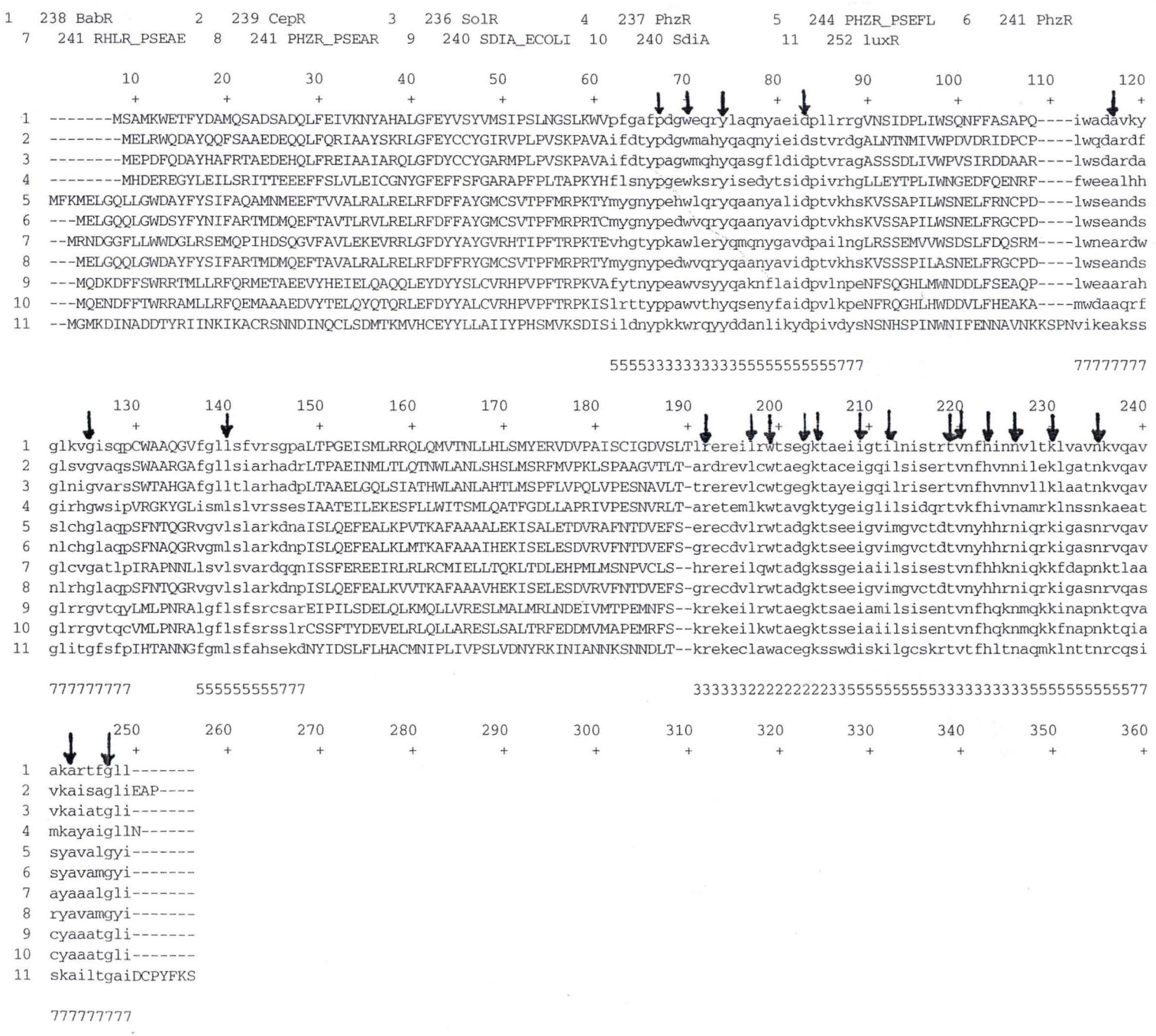


Figure 28 : Alignement Match-Box des séquences du second groupe. Les résidus conservés dans toutes les séquences sont marqués d'une flèche

calcul des méthodes employées plus tard et d'éviter les biais dans les résultats. On obtient un ensemble de 33 homologues appartenant à la famille LuxR, LuxR se retrouvant bien parmi ceux-ci.

Nous nous intéresserons plus loin à certaines caractéristiques structurales des homologues les plus proches de BabR ; c'est pourquoi, nous avons encore subdivisé le set en trois groupes de séquences:
- le groupe des 9 homologues dont l'identité avec BabR est supérieure à 30%
- le groupe des 9 homologues dont l'identité avec BabR est supérieure à 30% + LuxR
- le set en entier (33 séquences)

L'intérêt du second groupe par rapport au premier est le fait qu'il contient LuxR (25% d'identité avec BabR d'après l'alignement donné par ALIGN) et donc les informations dont nous disposons à son sujet pourront éventuellement être extrapolées aux autres séquences du groupe à partir d'un alignement multiple de ces séquences. Ainsi, l'idéal serait de pouvoir utiliser le second groupe dans les étapes suivantes. Seulement, l'ajout de LuxR qui n'est pas un homologue proche de BabR (<30% d'identité), pourrait perturber l'alignement par rapport à celui donné pour les séquences du premier groupe. Nous vérifierons cela dans le point 2. Nous comparerons aussi les alignements des trois groupes pour déterminer les régions les plus conservées.

Une autre classification pourrait donner d'autres informations. C'est pourquoi, nous avons regroupé les 33 homologues selon le type d'HSL qu'ils lient. D'après les données bibliographiques dont nous disposions, nous avons pu classer 19 homologues :

type d'HSL	**Les régulateurs correspondants**
C8HSL	CepR, SolR, YtbR
C6HSL	PhzR psear, YenR, PhzR pseae, YpsR
C4HSL	RhlR pseae, AhyR
7cis-C14HSL	CerR
3-oxo-C6HSL	LuxR, EsaR, ExpR, CarR, YpsR
3-oxo-C10HSL	VanR
3-oxo-C12HSL	LasR
3-oxo-C8HSL	TraR
3-OH-7cis-C14HSL	RhiR

Nous observons que CepR et SolR, les deux homologues les plus similaires à BabR, lient le même type de phéromone. La phéromone identifiée chez *B.abortus* ressemble à celle qui lie CepR et SolR puisque toutes les deux sont non-substituées et leur chaîne acyl a presque la même longueur (8 C pour celle se liant à CepR et SolR et 12 C pour celle trouvée chez *B.abortus*). Ceci est un argument pour dire que la dDHL trouvée chez *B.abortus* est bien la phéromone qui se lie à BabR.

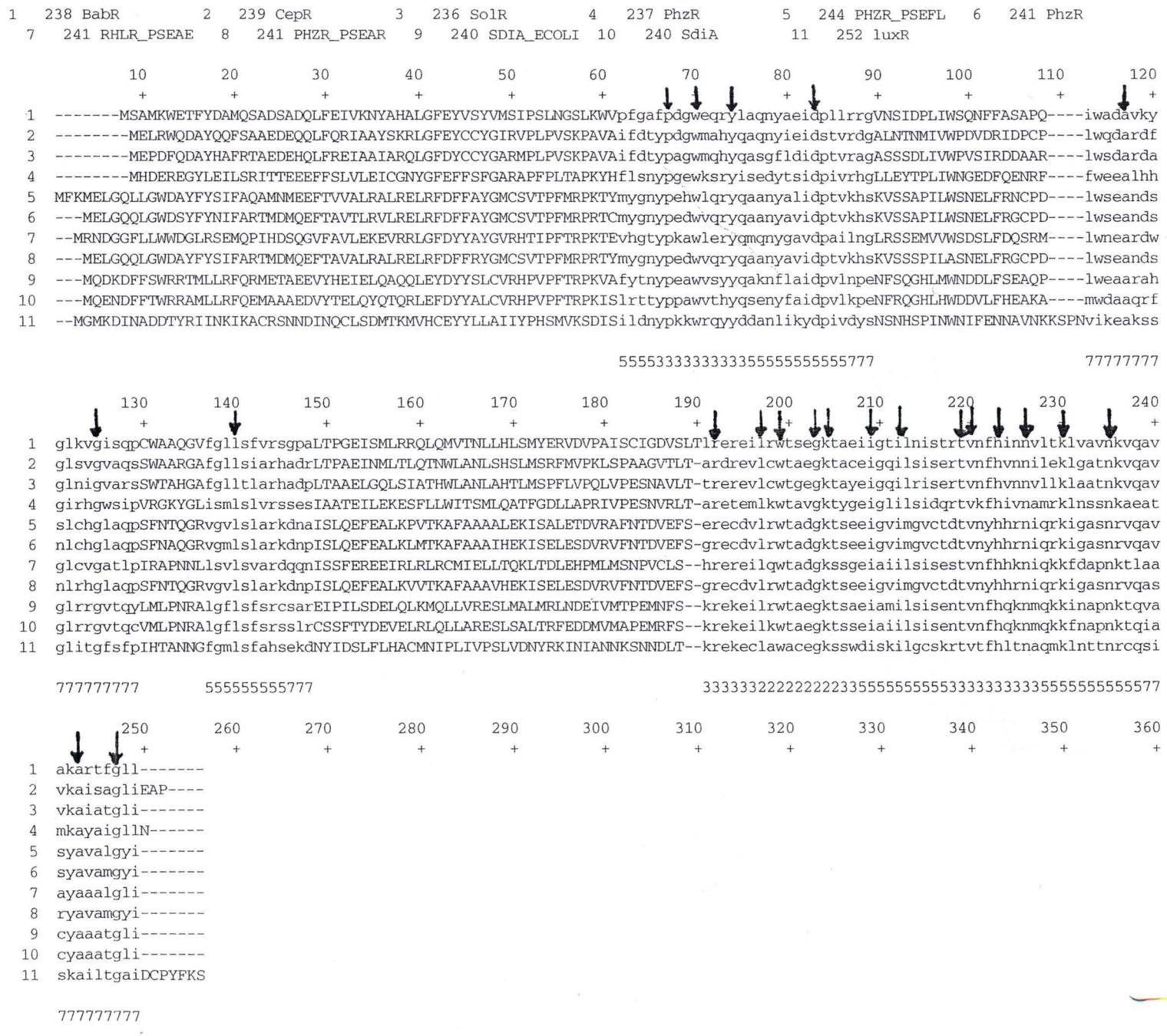


Figure 28 : Alignement Match-Box des séquences du second groupe. Les résidus conservés dans toutes les séquences sont marqués d'une flèche

```
                    ClustalW Multiple Sequence Alignment Results

                              Courtesy of the BCM Search Launcher


Page 1.1
                  1           15 16           30 31           45 46           60 61           75 76           90
    1 CepR        --------MELRWQD AYQQFSAAEDEQQLF QRIAAYSKRLGFEYC CYGIRVPLPVSKPAV AIFDTYPDGWMAHYQ AQNYIEIDSTVRDGA     82
    2 SolR        --------MEPDFQD AYHAFRTAEDEHQLF REIAAIARQLGFDYC CYGARMPLPVSKPAV AIFDTYPAGWMQHYQ ASGFLDIDPTVRAGA     82
    3 BabR        -------MSAMKWET FYDAMQSADSADQLF EIVKNYAHALGFEYV SYVMSIPSLNGSLKW VPFGAFPDGWEQRYL AQNYAEIDPLLRRGV     83
    4 PhzRpsech   ---MELGQQLGWDSY FYNIFARTMDMQEFT AVTLRVLRELRFDFF AYGMCSVTPFMRPRT CMYGNYPEDWVQRYQ AANYAVIDPTVKHSK     87
    5 PHZR_PSEAR  ---MELGQQLGWDAY FYSIFARTMDMQEFT AVALRALRELRFDFF RYGMCSVTPFMRPRT YMYGNYPEDWVQRYQ AANYAVIDPTVKHSK     87
    6 PHZR_PSEFL  MFKMELGQLLGWDAY FYSIFAQAMNMEEFT VVALRALRELRFDFF AYGMCSVTPFMRPKT YMYGNYPEHWLQRYQ AANYALIDPTVKHSK     90
    7 SDIA_ECOLI  ---MQDKDFFSWRRT MLLRFQRMETAEEVY HEIELQAQQLEYDYY SLCVRHPVPFTRPKV AFYTNYPEAWVSYYQ AKNFLAIDPVLNPEN     87
    8 SdiA        ---MQENDFFTWRRA MLLRFQEMAAAEDVY TELQYQTQRLEFDYY ALCVRHPVPFTRPKI SLRTTYPPAWVTHYQ SENYFAIDPVLKPEN     87
    9 RHLR_PSEAE  --MRNDGGFLLWWDG LRSEMQPIHDSQGVF AVLEKEVRRLGFDYY AYGVRHTIPFTRPKT EVHGTYPKAWLERYQ MQNYGAVDPAILNGL     88
   10 PhzRpseae   --------MHDEREG YLEILSRITTEEEFF SLVLEICGNYGFEFF SFGARAPFPLTAPKY HFLSNYPGEWKSRYI SEDYTSIDPIVRHGL     82
   11 luxR        --MGMKDINADDTYR IINKIKACRSNNDIN QCLSDMTKMVHCEYY LLAIIYPHSMVKSDI SILDNYPKKWRQYYD DANLIKYDPIVDYSN     88


Page 2.1
                  91          105 106         120 121         135 136         150 151         165 166         180
    1 CepR        LNTNMIVWPDVD--- -RIDPCPLWQDARDF GLSVGVAQSSWAARG AFGLLSIARHADRLT PAEINMLTLQTNWLA NLSHS-LMSRFMVPK    167
    2 SolR        SSSDLIVWPVSI--- -RDDAARLWSDARDA GLNIGVARSSWTAHG AFGLLTLARHADPLT AAELGQLSIATHWLA NLAHT-LMSPFLVPQ    167
    3 BabR        NSIDPLIWSQNF--- -FASAPQIWADAVKY GLKVGISQPCWAAQG VFGLLSFVRSGPALT PGEISMLRRQLQMVT NLLHLSMYERVDVPA    169
    4 PhzRpsech   VSSAPILWSNEL--- -FRGCPDLWSEANDS NLCHGLAQPSFNAQG RVGMLSLARKDNPIS LQEFEALKLMTKAFA AAIHE-KISELESDV    172
    5 PHZR_PSEAR  VSSSPILASNEL--- -FRGCPDLWSEANDS NLRHGLAQPSFNTQG RVGVLSLARKDNPIS LQEFEALKVVTKAFA AAVHE-KISELESDV    172
    6 PHZR_PSEFL  VSSAPILWSNEL--- -FRNCPDLWSEANDS SLCHGLAQPSFNTQG RVGVLSLARKDNAIS LQEFEALKPVTKAFA AAALE-KISALETDV    175
    7 SDIA_ECOLI  FSQGHLMWNDDL--- -FSEAQPLWEAARAH GLRRGVTQYLMLPNR ALGFLSFSRCSAREI PILSDELQLKMQLLV RESLM-ALMRLNDEI    172
    8 SdiA        FRQGHLHWDDVL--- -FHEAKAMWDAAQRF GLRRGVTQCVMLPNR ALGFLSFSRSSLRCS SFTYDEVELRLQLLA RESLS-ALTRFEDDM    172
    9 RHLR_PSEAE  RSSEMVVWSDSL--- -FDQSRMLWNEARDW GLCVGATLPIRAPNN LLSVLSVARDQQNIS SFEREEIRLRLRCMI ELLTQ-KLTDLEHPM    173
   10 PhzRpseae   LEYTPLIWNGED--- -FQENRFFWEEALHH GIRHGWSIPVRGKYG LISMLSLVRSSESIA ATEILEKESFLLWIT SMLQA-TFGDLLAPR    167
   11 luxR        SNHSPINWNIFENNA VNKKSPNVIKEAKSS GLITGFSFPIHTANN GFGMLSFAHSEKDNY IDSLFLHACMNIPLI VPSLVDNYRKINIAN    178


Page 3.1
                  181         195 196         210 211         225 226         240 241         255 256         270
    1 CepR        LSPAAGVTLTARDRE VLCWTAEGKTACEIG QILSISERTVNFHVN NILEKLGATNKVQAV VKAISAGLIEAP--- -    239
    2 SolR        LVPESNAVLTTRERE VLCWTGEGKTAYEIG QILRISERTVNFHVN NVLLKLAATNKVQAV VKAIATGLI------ -    236
    3 BabR        ISCIGDVSLTLRERE ILRWTSEGKTAEIIG TILNISTRTVNFHIN NVLTKLVAVNKVQAV AKARTFGLL------ -    238
    4 PhzRpsech   RVFNTDVEFSGRECD VLRWTADGKTSEEIG VIMGVCTDTVNYHHR NIQRKIGASNRVQAV SYAVAMGYI------ -    241
    5 PHZR_PSEAR  RVFNTDVEFSGRECD VLRWTADGKTSEEIG VIMGVCTDTVNYHHR NIQRKIGASNRVQAS RYAVAMGYI------ -    241
    6 PHZR_PSEFL  RAFNTDVEFSERECD VLRWTADGKTSEEIG VIMGVCTDTVNYHHR NIQRKIGASNRVQAV SYAVALGYI------ -    244
    7 SDIA_ECOLI  VMTP-EMNFSKREKE ILRWTAEGKTSAEIA MILSISENTVNFHQK NMQKKINAPNKTQVA CYAAATGLI------ -    240
    8 SdiA        VMAP-EMRFSKREKE ILKWTAEGKTSSEIA IILSISENTVNFHQK NMQKKFNAPNKTQIA CYAAATGLI------ -    240
    9 RHLR_PSEAE  LMSN-PVCLSHRERE ILQWTADGKSSGEIA IILSISESTVNFHHK NIQKKFDAPNKTLAA AYAAALGLI------ -    241
   10 PhzRpseae   IVPESNVRLTARETE MLKWTAVGKTYGEIG LILSIDQRTVKFHIV NAMRKLNSSNKAEAT MKAYAIGLLN----- -    237
   11 luxR        NKSN--NDLTKREKE CLAWACEGKSSWDIS KILGCSKRTVTFHLT NAQMKLNTTNRCQSI SKAILTGAIDCPYFK S    252
```

Figure 29 : Alignement ClustalW des séquences du second groupe

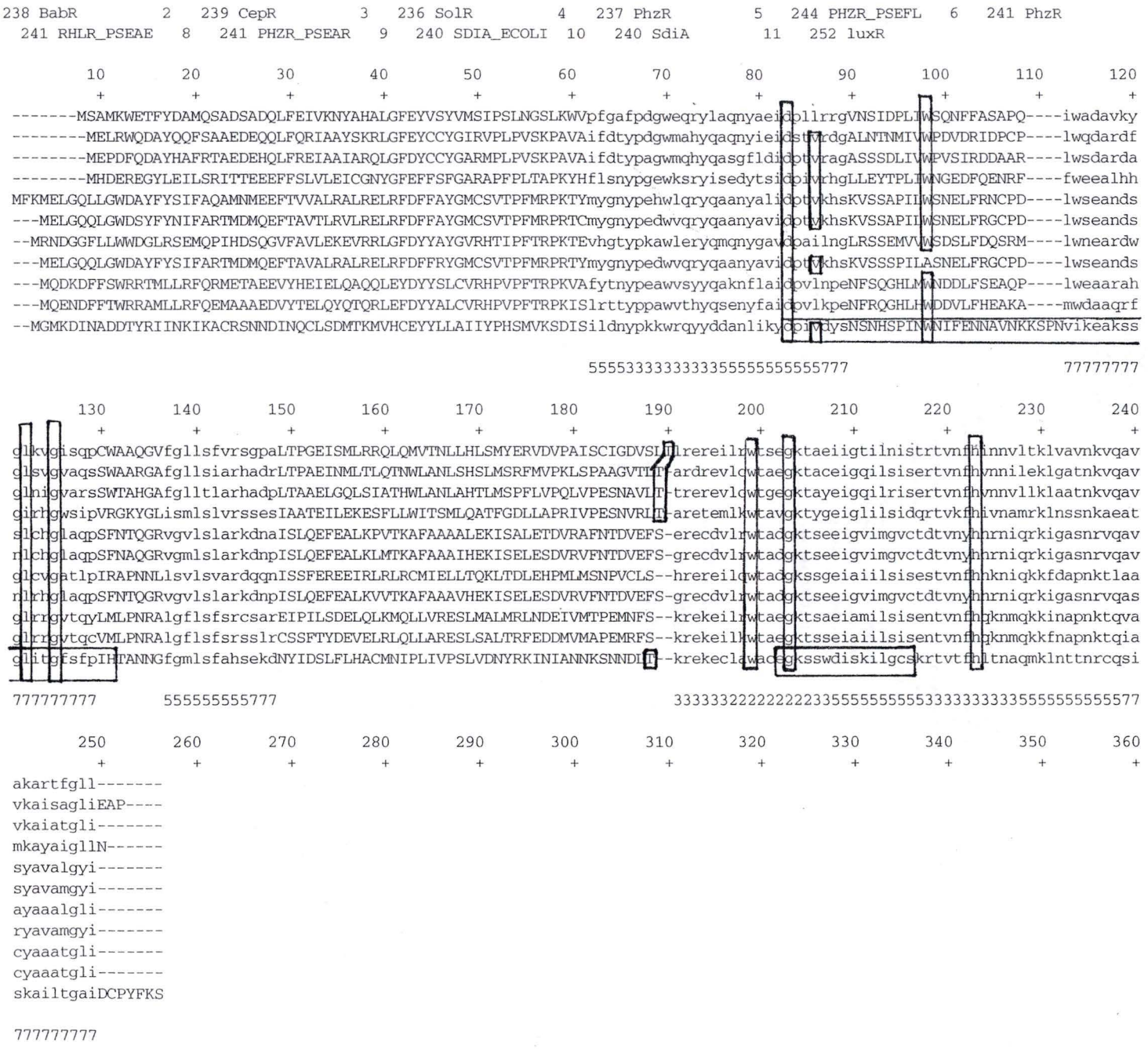


Figure 30 : La première région encadrée correspond au domaine de liaison à la phéromone et la deuxième au domaine de liaison à l'ADN. Les résidus encadrés correspondent aux résidus ayant été prédits comme importants dans LuxR par des expériences de mutagenèse dirigée ET qui sont conservés parmi les homologues

Nous poursuivons notre analyse dans l'étape suivante avec les trois groupes définis par la première classification. Nous allons donc pouvoir vérifier si l'utilisation du second groupe incluant LuxR pourrait être envisagée lors des prédictions structurales, sans grand risque d'erreurs causées par un mauvais alignement. De plus, la comparaison des trois groupes de séquences permettra de déterminer les régions les plus conservées et donc les plus importantes d'un point de vue fonctionnel.

2. Alignements multiples

Nous avons soumis les trois groupes, définis précédemment, aux algorithmes de Match-Box et de ClustalW. La première observation que nous pouvons faire est que les deux plus grandes boîtes données par Match-Box sont conservées dans les trois groupes. Elles correspondent aux boîtes possédant le plus de résidus conservés et dont les scores sont les plus significatifs (figure 28). Les autres boîtes, obtenues avec le set minimal de séquences, ont tendance à disparaître quand le set de séquences s'agrandit. Les deux boîtes principales ainsi qu'une autre boîte intermédiaire importante, résultats de l'alignement des séquences du premier groupe, se retrouvent dans l'alignement des séquences du deuxième groupe. L'ajout de LuxR n'a donc pas trop modifié l'alignement des séquences du groupe 1. L'alignement étant assez plausible, nous pourrons utiliser le groupe le plus intéressant, c'est-à-dire le groupe 2, dans les opérations suivantes ; les deux autres ne nous serviront plus.

Les alignements obtenus par ClustalW sont comparables à ceux de Match-Box (figure 29). Les régions importantes sont alignées de la même manière par les deux programmes. Nous pouvons donc accorder une confiance assez élevée à ces alignements.

Si nous prenons en compte les informations que l'on possède sur LuxR, nous pouvons déduire, à partir de l'alignement des séquences du groupe 2, que les deux boîtes principales seraient attachées aux régions de liaison à la phéromone et de liaison à l'ADN. De plus, certains résidus essentiels à l'activité de LuxR[1] sont fortement conservés. Ceux-ci jouent donc probablement un rôle important dans les autres protéines de la famille aussi (figure 30) :

- l'aspartate 79, la valine 82 , le tryptophane 94, la leucine 118 et la glycine 121 sont localisés dans la **région de liaison à la phéromone** ;

[1] Ce sont des expériences de mutagenèse dirigée qui ont permis de déterminer l'importance de ces résidus (cfr figure 17).

Scores attribués aux différentes méthodes de prédictions de structures secondaires.

	Hélices α		Brins β	
Méthode	Niveau de confiance	Score	Niveau de confiance	Score
PHD	8-9	3	>5	-3
	5-7	2	3-4	-2
	<5	1	<3	-1
PSIPRED	8-9	3	>5	-3
	5-7	2	3-4	-2
	<5	1	<3	-1
PROF	Entre les extrêmes	2	Entre les extrêmes	-2
	Extrémités	1	Extrémités	-1
PREDATOR/JPRED		1		1
Seuil après total	Confiant	≥7	Confiant	≤-5
	confiance	6	confiance	-4
	moindre		moindre	

Tableau 8

- la thréonine 184, le tryptophane 193, la glycine 197 et l'histidine 217 sont localisés dans la **région de liaison à l'ADN**.

D'autres résidus sont fortement conservés ; leur importance n'est pas à négliger. Toutes ces données nous permettent de dire que deux régions, l'une liant la phéromone et l'autre l'ADN, apparaissent dans l'alignement pour chaque protéine du second groupe.

Une autre information peut être tirée de l'alignement des séquences du second set et des données que l'on possède sur LuxR : nous savons que la région limitant les deux domaines de LuxR est centrée sur le résidu 160. Cette position dans l'alignement détermine ainsi la limite potentielle des deux domaines dans les autres protéines.

Nous allons donc nous tourner vers la prédiction de structures secondaires. Ainsi, des données structurales vont venir compléter les informations obtenues par les alignements.

3.Prédictions de structures secondaires

BabR et les 10 homologues du groupe 2 ont été soumis à cinq méthodes de prédictions de structures secondaires décrites dans le chapitre « Matériel et Méthodes » : PHD, PROF, PSIpred, PREDATOR et JPRED2. Chaque séquence a été envoyée séparément sur le site Internet des méthodes PROF, PSIpred et JPRED2. Par contre, en ce qui concerne les méthodes pour lesquelles il est possible et recommandé d'envoyer un set de séquences homologues, c'est-à-dire PHD et PREDATOR, nous avons envoyé sur leur site l'ensemble des 11 séquences. Les résultats sont présentés dans l'annexe 1. Les prédictions de PHD et PSIpred sont accompagnées d'un score de confiance. Plus le score (de 1 à 9) est haut et plus la prédiction est certaine.

Afin d'obtenir des prédictions plausibles pour chaque séquence, nous avons construit un consensus des méthodes citées ci-dessus. En outre, celui-ci permet de mieux distinguer la succession des structures secondaires. Il a été réalisé selon une méthodologie déjà employée au laboratoire (de Fays *et al.*, 1999). Les prédictions de l' annexe 1 sont remplacées par des scores définis dans le tableau 8. Un poids plus important a été attribué aux deux méthodes les plus plausibles c'est-à-dire PSIpred et PHD. PHD est une méthode renommée et PSIpred a fourni les meilleurs résultats au CASP3. Les trois autres méthodes pourraient donner de meilleurs résultats que PHD mais elles doivent encore faire leurs preuves. C'est pourquoi nous leur donnons un poids plus faible. En général les brins sont moins bien prédits que les hélices et donc le seuil d'attribution des scores a été abaissé par rapport à celui des hélices. Un total est ensuite calculé pour chaque résidu. Deux niveaux de confiance sont proposés. Si

	Topologie prédite
BabR	ααββααβαββαααααα
CepR	αββαβαβαααααα
SolR	ααββαβαββαααααα
PhzR Pseae	ααββαβαβαααααα
PhzR Psefl	ααββααβαββαααααα
PhzR Psech	ααββααβαβαααααα
RhlR Pseae	βααββααβαβαααααα
PhzR Psear	ααββαβαββαααααα
SdiA E.Coli	ααβααβαβαααααα
SdiA Salmonella	αβααβαβαααααα
LuxR	αααβαβαββαααααα

Tableau 9 : synthèse des topologies prédites.

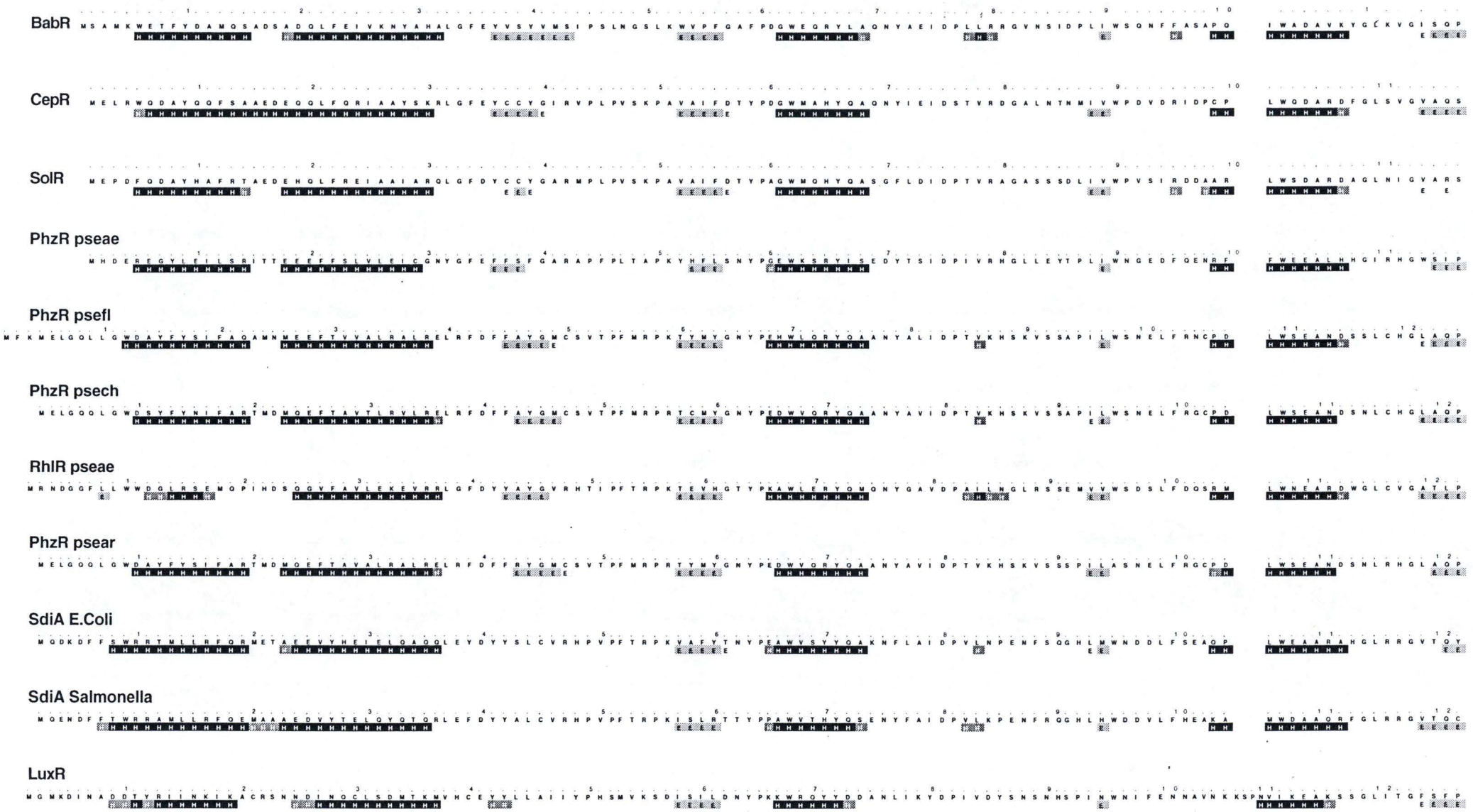


Fig 31 a : alignement des domaines N-terminaux des homologues de BabR

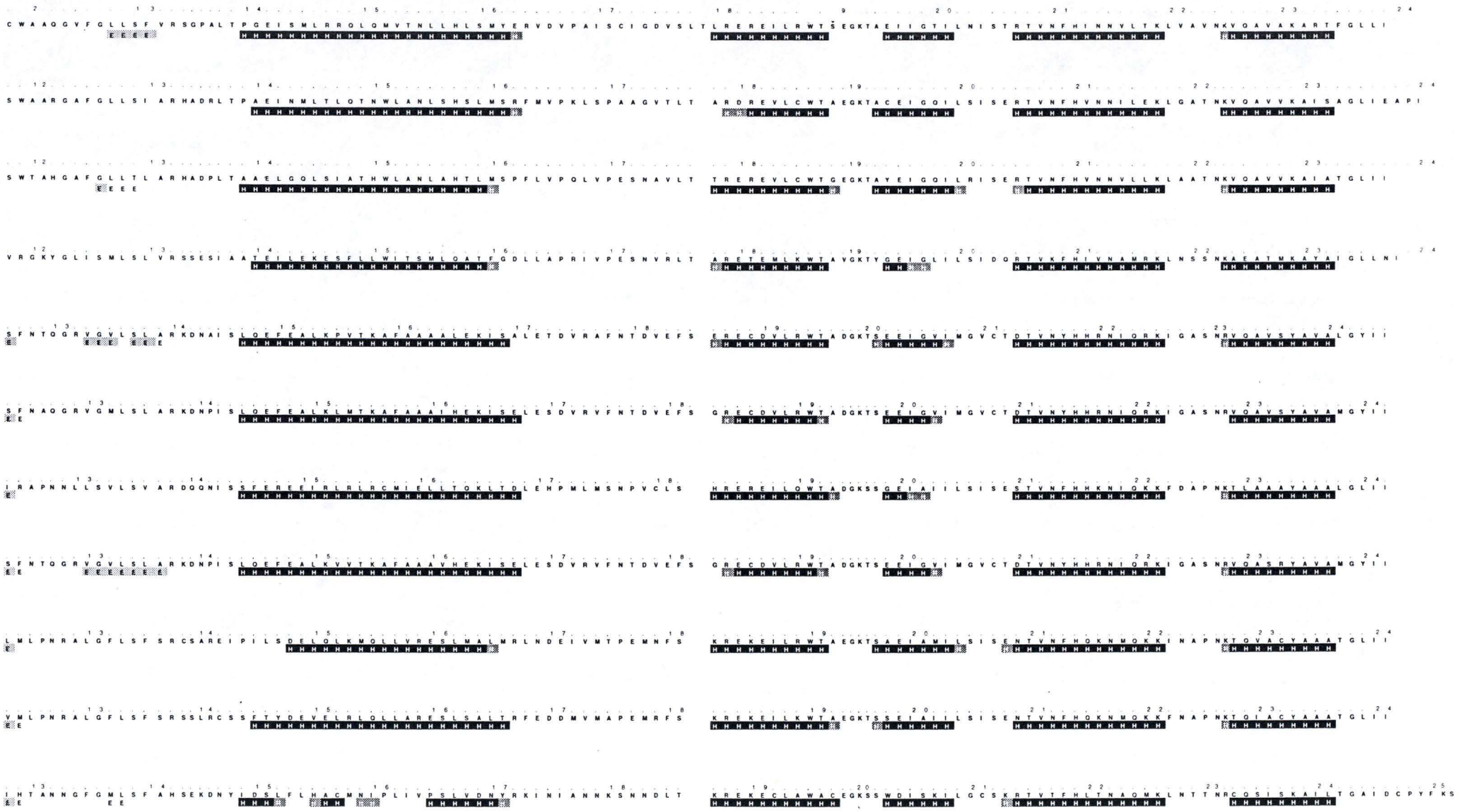


Fig 31 b : alignement des domaines C-terminaux des homologues de BabR

le total est supérieur ou égal à 7, le résidu est assigné à une hélice α et s'il est inférieur ou égal à –5, le résidu est assigné à un brin β. Les scores 6 et –4 définissent une hélice et un brin respectivement, mais de moindre confiance. Les consensus finaux sont rassemblés dans la figure 31. Nous pouvons observer que Match-Box aligne les structures secondaires prédites : la confiance dans l'alignement de séquences et dans les prédictions de structures secondaires est corrélée. Une synthèse de ces prédictions est présentée dans le tableau 9.

D'après ce que nous avons vu dans l'étape d'alignement et si nous comparons cela aux structures secondaires prédites, la région limitant les deux domaines des protéines correspondrait environ, dans chaque protéine, à la cinquième hélice α en commençant par la fin. Ainsi, cette hélice pourrait jouer le rôle d'un lien entre les deux domaines. Nous pouvons remarquer que les domaines N-terminaux des protéines sont de type α/β et qu'ils varient légèrement du point de vue de la topologie, alors que les domaines C-terminaux sont toujours composés de 4 hélices α. Ceci paraît logique puisque les domaines N-terminaux sont impliqués dans la liaison à des phéromones de longueur et de biochimie variables alors que les domaines C-terminaux lient toujours le même substrat c'est-à-dire l'ADN *via* leur motif HTH.

Nous allons nous intéresser plus en détail à BabR dans les étapes de prédictions de structures tertiaires en traitant les deux domaines de BabR séparément. En effet, les deux domaines adoptent probablement deux « folds » distincts et ont deux fonctions différentes : leur séparation ne peut que diminuer le bruit de fond lors de recherches en banques de structures tertiaires. Pour éviter de couper un fragment d'un domaine ou de l'autre, nous avons scindé BabR en deux parties de telle sorte qu'une zone de recouvrement appartienne aux deux domaines. Cette zone de recouvrement correspond environ à la cinquième hélice α en commençant par l'extrémité C-terminale.

Au terme de cette opération nous avons donc pu déterminer la topologie de BabR et de ses 9 homologues les plus proches ainsi que celle de LuxR. En outre, les protéines ont pu être divisées d'une façon approximative en deux domaines. Notre approche plus détaillée de BabR va se poursuivre sur chacun des domaines pris séparément.

Tableau 10

THREADER 2.5

domaine C-terminal

			CLASSIFICATION CATH			
	code PDB	**Z-score**	**class**	**architecture**	**topology**	**homologous superfamily (PDB code)**
1	1pda 03	2.68	Alpha Beta	2-Layer Sandwich	Human Macrophage Inflammatory Protein 1 Beta, subunit A	1pda domain 3
2	1abv 00	2.34	Mainly Alpha	Non-Bundle	Peroxidase, domain 1	1abv
3	1srr A0	2.21	Alpha Beta	3-Layer(aba) Sandwich	Rossmann fold (Nitrogenase Molybdenum-Iron Protein, subunit A, domain 3)	1efu chain A domain 1
4	1kpt A0	2.20	Alpha Beta	2-Layer Sandwich	Killer Toxin P4, subunit A	1kpt chain A
5	1bnc A1	2.18	Alpha Beta	3-Layer(aba) Sandwich	Rossmann fold (Nitrogenase Molybdenum-Iron Protein, subunit A, domain 3)	1iow domain 1

Tableau 11

3DPSSM

domaine C-terminal

			CLASSIFICATION CATH			
	code PDB	**E-value**	**class**	**architecture**	**topology**	**homologous superfamily (PDB code)**
1	1rnl _1	2.71e-03	Mainly Alpha	Non-Bundle	Arc Repressor Mutant, subunit A	1mbe
2	2trt _1	1.81e-02	Mainly Alpha	Non-Bundle	Tetracycline Repressor, domain 2	2tct domain 2
3	1rnl __	5.03e-02	Mainly Alpha	Non-Bundle	Arc Repressor Mutant, subunit A	1mbe
4	2wrp R0	3.72e-01	Mainly Alpha	Non-Bundle	Arc Repressor Mutant, subunit A	1mbe
5	1vol A2	9.26e-01	Mainly Alpha	Non-Bundle	Cyclin A, domain 1	1vin domain 1

GENTHREADER Tableau 12

domaine C-terminal

			CLASSIFICATION CATH			
	code PDB	**probabilité**	**class**	**architecture**	**topology**	**homologous superfamily (PDB code)**
1	1a04 A0	1	Mainly Alpha	Non-Bundle	Arc Repressor Mutant, subunit A	1mbe
2	1pdn C0	1	Mainly Alpha	Non-Bundle	Arc Repressor Mutant, subunit A	1lea
3	1smt A0	0,533	Mainly Alpha	Non-Bundle	Arc Repressor Mutant, subunit A	1lea
4	1nct 00	0,425	Mainly Beta	Sandwich	Immunoglobulin-like	1zxq domain 2
5	1r69 00	0,169	Mainly Alpha	Non-Bundle	434 Repressor (Amino-terminal Domain)	1neq

Tableau 13

TOPITS

domaine C-terminal

			CLASSIFICATION CATH			
	code PDB	**Z-score**	**class**	**architecture**	**topology**	**homologous superfamily (PDB code)**
1	1a04 _A	2.21	Mainly Alpha	Non-Bundle	Arc Repressor Mutant, subunit A	1mbe
2	1imb _A	1.89	Alpha Beta	2-Layer Sandwich	Fructose-1,6-Bisphosphatase, subunit A, domain 1	2hhm chain A domain 1
3	1asz _B	1.87	Alpha Beta	3-Layer(aba) Sandwich	Aspartyl tRNA Synthetase, subunit A, domain 2	1asy chain A domain 2
4	1a34 _A	1.77	Mainly Beta	Sandwich	Satellite Panicum Mosaic Virus, chain A	1stm chain A
5	1ayr _A	1.76	/	/	/	/

Tableau 14

PSI-BLAST BORK

domaine C-terminal

		CLASSIFICATION CATH			
code PDB	**E-value**	**class**	**architecture**	**topology**	**homologous superfamily (PDB code)**
1a04 _A	8e-14	Mainly Alpha	Non-Bundle	Arc Repressor Mutant, subunit A	1mbe

4.Le domaine C-terminal de BabR

4.1. La modélisation

4.1.1.Prédictions de la structure tertiaire

Nous avons commencé par essayer de savoir si nous pouvions déterminer la structure tertiaire du domaine C-terminal de BabR par homologie. Pour ce faire, une recherche a été entreprise dans la banque PDB à l'aide du programme BLASTP. Un seul homologue a été retrouvé, ce qui suffit pour pouvoir réaliser une modélisation par homologie ; son code d'accession dans la banque PDB est 1a04 (Baikalov *et al.*, 1998). NarL, puisqu'il s'agit de lui, est un régulateur de la transcription d'*E.coli*, intervenant dans un système à deux composants (Baikalov *et al.*, 1996). Il manque, dans la forme cristalline, les coordonnées des résidus 150 à 154. Ceci est dû à la flexibilité de ce lien entre les deux domaines de NarL ; son instabilité empêche d'avoir une image correcte de cette portion en 3D. NarL a été cristallisé sous forme d'un dimère ; en effet, il est actif sous cette forme. NarL existe également dans la banque PDB sous le code d'accession 1rnl : il a cette fois été cristallisé sous forme monomérique et la portion manquante dans la structure est comprise entre les résidus 143 et 154 (Baikalov *et al.*, 1996). Cette entrée de PDB est antécédente à l'entrée 1a04 et sa résolution est moins bonne que celle de 1a04[1]. NarL n'a pas été cristallisé sous forme d'un complexe avec l'ADN.

Nous utiliserons donc bien dans ce cas la stratégie de modélisation par homologie. Cependant, pour obtenir des alignements séquence-structure en plus des alignements séquence-séquence et pour avoir des confirmations supplémentaires de ce patron, nous avons prospecté la prédiction de « folds » par différentes méthodes. Ces alignements séquence-structure permettront d'optimaliser d'avantage l'alignement séquence-structure final. Nous avons utilisé THREADER 2.5, une méthode de « threading » ainsi que GENTHREADER, 3DPSSM et TOPITS, trois méthodes de « pseudo-threading ». En dehors de ces deux catégories, PSI-BLAST BORK a également servi de moyen de prédiction de la structure tertiaire. Les tableaux 10 à 13 reprennent les cinq meilleurs « hits » de chaque méthode. Le tableau 14 correspond au résultat donné par PSI-BLAST BORK. Nous observons que 4 méthodes sur 5 placent le domaine C-terminal de NarL en tête de classement. Ceci confirme clairement ce que nous avions obtenu par BLASTP c'est-à-dire que NarL est le patron idéal pour modéliser le domaine C-terminal de BabR par homologie.

Maintenant que nous avons pu déterminer la protéine de structure tridimensionnelle connue la mieux adaptée au domaine C-terminal de BabR, nous

[1] 2,2 Å pour 1a04 contre 2,4 Å pour 1rnl

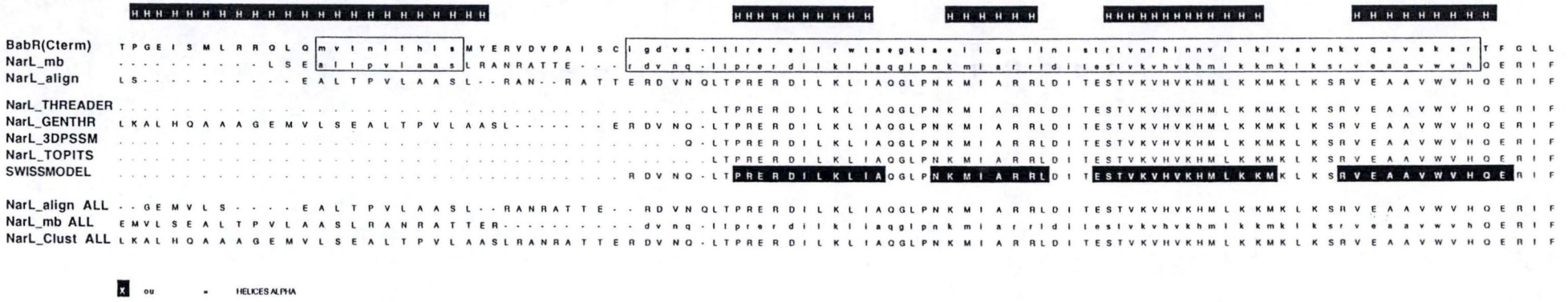


Figure 32 : alignements séquence-structure

envisageons ci-dessous l'alignement séquence-structure optimal essentiel à l'obtention d'un bon modèle.

4.1.2.L'alignement séquence-structure

Si nous voulons parvenir à un modèle relativement correct, la séquence doit être alignée à la structure de la façon la plus précise possible. Pour ce faire, nous avons décidé de comparer les alignements de plusieurs méthodes et d'en tirer un alignement consensus. Les résultats des méthodes d'alignements séquence-séquence utilisées précédemment, c'est-à-dire Match-Box, ClustalW et ALIGN, ont également été pris en compte. Ces différents alignements sont présentés à la figure 32. Nous remarquons que les 61 derniers acides aminés de BabR sont alignés à NarL de la même façon par toutes les méthodes. La zone manquante du fichier PDB de NarL, que l'on appelle « gap de structure », est localisée juste en amont de cette région. Cette zone, ainsi que l'hélice α n°6 adjacente[2], constitueraient le lien entre les deux domaines de BabR. En effet, l'alignement de cette région avec NarL diffère d'une méthode à l'autre, démontrant son caractère variable. Ce dernier est caractéristique d'une région probablement beaucoup moins importante fonctionnellement et donc plus sujette à diverses mutations, telle qu'une région de liaison entre deux domaines.

Nous pouvons conclure que les 61 derniers acides aminés de BabR, correspondant « exactement » au domaine C-terminal, seront modélisés d'une façon fiable puisque l'alignement de ceux-ci avec NarL est lui aussi très crédible. La modélisation des 39 autres résidus du domaine C-terminal, c'est-à-dire celle du lien, pose plus problème. Aucun des alignements donnés ci-dessus ne permet de dire quel serait l'alignement optimal. C'est pourquoi, nous avons décidé d'aligner cette portion de la manière la plus simple qui soit puisque les deux séquences ont environ la même longueur, c'est-à-dire en plaçant les 61 derniers acides aminés à la suite des 39 premiers sans introduire de « gaps », ni dans NarL ni dans le domaine C-terminal de BabR. Rendue plausible par la minimisation d'énergie, cette zone du modèle sera cependant incertaine.

Nous avons déterminé ici un alignement séquence-structure qui sera utilisé dans l'étape suivante pour construire le modèle du domaine C-terminal de BabR. Ce modèle pourra être considéré comme correct pour les 61 derniers résidus c'est-à-dire le domaine C-terminal de BabR à proprement parler.

4.1.3.Le modèle

Nous avons utilisé le programme MODELLER4 pour cette dernière étape de la modélisation du domaine C-terminal de BabR, c'est-à-dire l'assignation des coordonnées. MODELLER4 requiert le fichier de coordonnées du patron NarL ainsi

[2] La cinquième en commençant par la fin

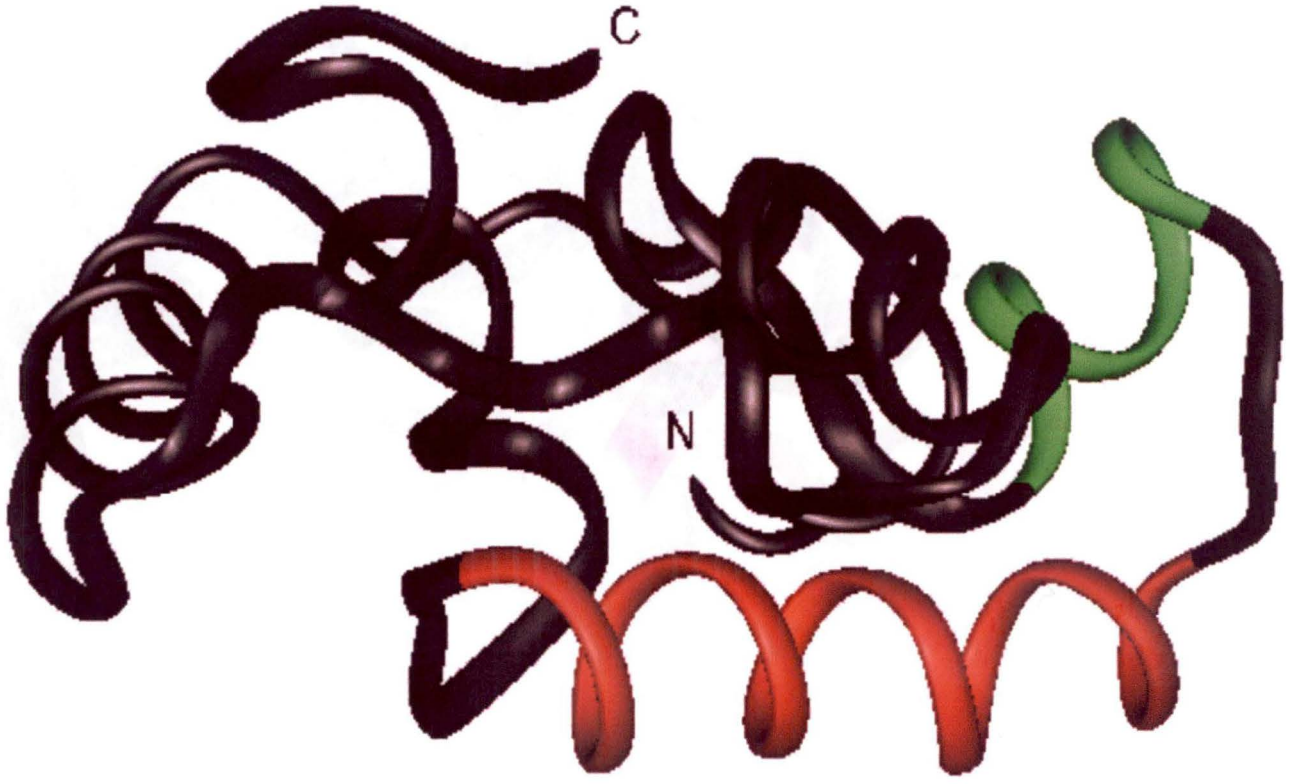


Fig. 33: Position des deux hélices du motif HTH dans le modèle tridimensionnel du domaine C-terminal de BabR. La première hélice du motif est colorée en vert. La seconde hélice du motif est colorée en rouge. Les lettres N et C indiquent les extrémités N- et C-terminales du domaine.

qu'un fichier « nom.ali » contenant l'alignement séquence-structure. Après exécution du programme, le modèle du domaine C-terminal de BabR a été soumis à DISCOVER afin de minimiser l'énergie du système et d'affiner ainsi le modèle. Le nombre d'itérations que doit effectuer le programme pour minimiser l'énergie a été fixé à 10000. A l'aide d'une option de DISCOVER, nous avons fixé le squelette des quatre dernières hélices de telle sorte que seules les chaînes latérales, la région incertaine et les boucles soient soumises à l'algorithme de mécanique moléculaire. Le modèle final est montré à la figure 33. Le domaine HTH est constitué des hélices α n° 8 et 9. Voici la topologie de BabR en entier et, en encadré, le domaine HTH : αααββαααβαββααααα.

Le modèle du domaine C-terminal de LuxR a été construit de la même manière que celui du domaine C-terminal de BabR :

- le même « fold » a été prédit par les méthodes de prédictions de structures tertiaires utilisées auparavant : NarL
- l'alignement séquence-structure a été déterminé de la même manière : la région incertaine, jouant le rôle d'un lien entre les deux domaines de LuxR, a été alignée de manière approximative à NarL.
- la minimisation d'énergie effectuée par DISCOVER a été effectuée en fixant les quatre dernières hélices. Le nombre d'itérations a également été fixé à 10000. Le modèle de LuxR est montré à la figure 34.
- le domaine HTH est constitué des hélices α n° 9 et 10. Voici la topologie de LuxR en entier et, en encadré, le domaine HTH : αααβαβαββαααααα.

Maintenant que nous sommes en possession des modèles des domaines C-terminaux de BabR et LuxR qui nous ont permis de localiser leur domaine HTH, nous pouvons aborder la problématique de l'interaction HTH-ADN.

4.2. L'analyse des interactions HTH-ADN

Nous ne connaissons pas la boîte d'ADN à laquelle se lie BabR donc nous ne pouvons pas analyser les interactions directes entre le domaine HTH de BabR et sa boîte d'ADN préférentielle. Cependant, grâce aux informations que l'on possède sur d'autres complexes HTH-ADN de structure tridimensionnelle résolue ainsi que sur LuxR et sa *lux*box, nous pourrions tirer quelques conclusions intéressantes sur le domaine HTH de BabR.

Une recherche par mots-clés nous a permis de trouver dans PDB des complexes HTH-ADN ; nous en sélectionnons quatre de telle sorte que ces complexes ne soient pas redondants. Les quatre protéines impliquées dans ces quatre complexes sont données ci-dessous :

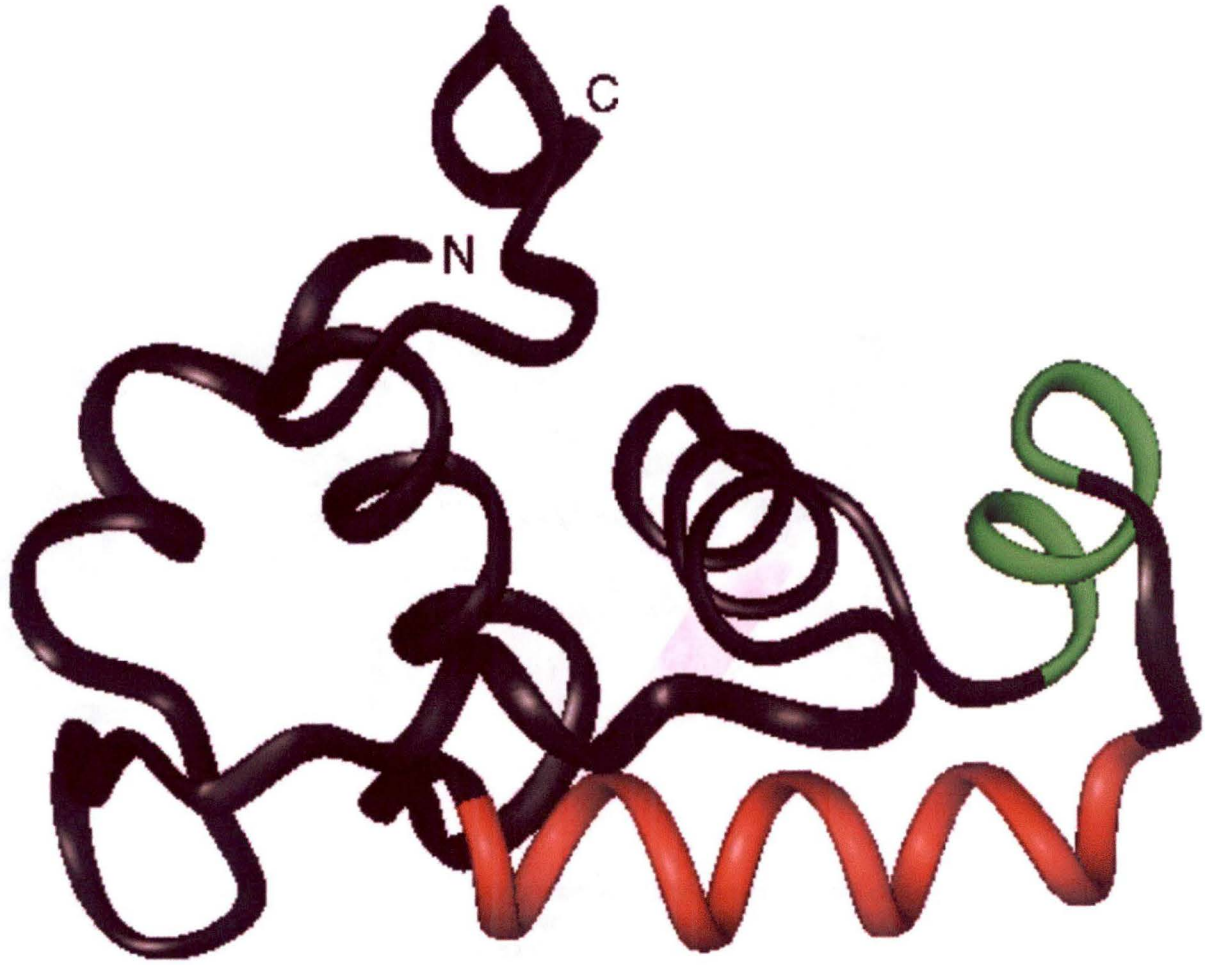
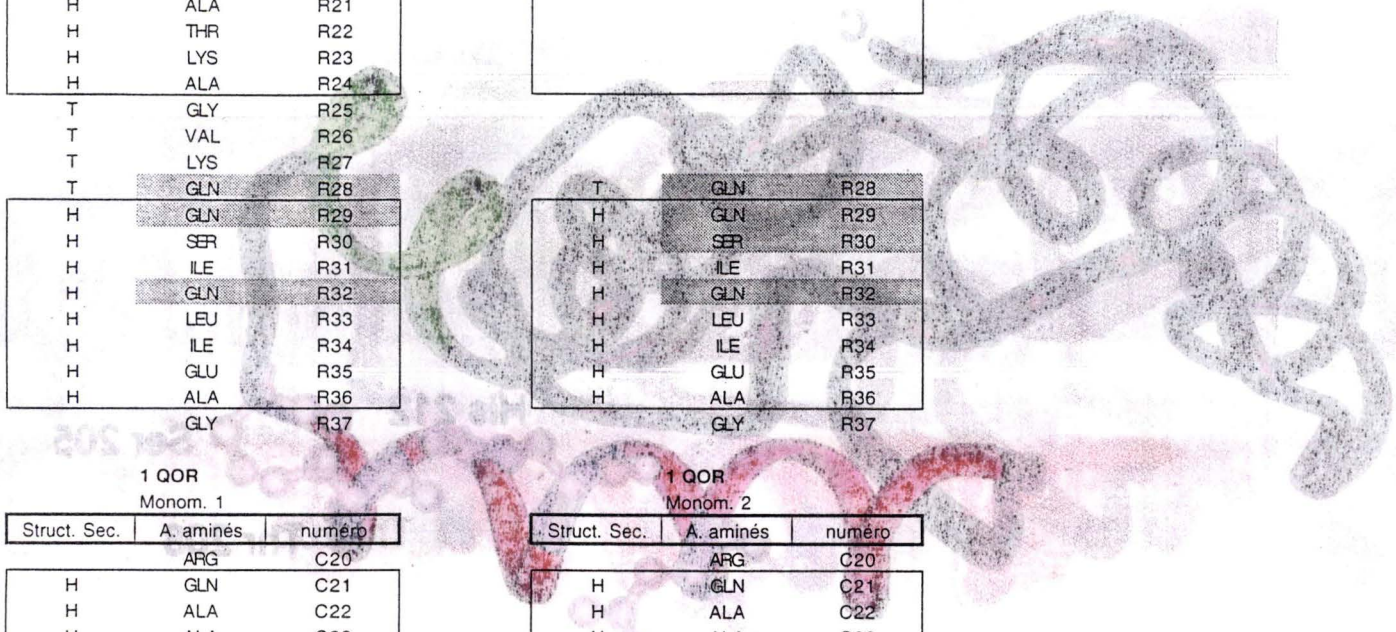


Fig. 34: Position des deux hélices du motif HTH dans le modèle tridimensionnel du domaine C-terminal de LuxR. La première hélice du motif est colorée en vert. La seconde hélice du motif est colorée en rouge. Les lettres N et C indiquent les extrémités N- et C-terminales du domaine.



3 CRO
Monom. 1

Struct. Sec.	A. aminés	numéro
	THR	R16
H	GLN	R17
H	THR	R18
H	GLU	R19
H	LEU	R20
H	ALA	R21
H	THR	R22
H	LYS	R23
H	ALA	R24
T	GLY	R25
T	VAL	R26
T	LYS	R27
T	GLN	R28
H	GLN	R29
H	SER	R30
H	ILE	R31
H	GLN	R32
H	LEU	R33
H	ILE	R34
H	GLU	R35
H	ALA	R36
	GLY	R37

3 CRO
Monom. 2

Struct. Sec.	A. aminés	numéro
T	GLN	R28
H	GLN	R29
H	SER	R30
H	ILE	R31
H	GLN	R32
H	LEU	R33
H	ILE	R34
H	GLU	R35
H	ALA	R36
	GLY	R37

1 QOR
Monom. 1

Struct. Sec.	A. aminés	numéro
	ARG	C20
H	GLN	C21
H	ALA	C22
H	ALA	C23
H	LEU	C24
H	GLY	C25
H	LYS	C26
H	MET	C27
H	VAL	C28
	GLY	C29
	VAL	C30
	SER	C31
	ASN	C32
H	VAL	C33
H	ALA	C34
H	ILE	C35
H	SER	C36
H	GLN	C37
H	TRP	C38
H	GLU	C39
H	ARG	C40
	SER	C41
	GLU	C42
	THR	C43

1 QOR
Monom. 2

Struct. Sec.	A. aminés	numéro
	ARG	C20
H	GLN	C21
H	ALA	C22
H	ALA	C23
H	LEU	C24
H	GLY	C25
H	LYS	C26
H	MET	C27
H	VAL	C28
	GLY	C29
	VAL	C30
	SER	C31
	ASN	C32
H	VAL	C33
H	ALA	C34
H	ILE	C35
H	SER	C36
H	GLN	C37
H	TRP	C38
H	GLU	C39
H	ARG	C40
	SER	C41
	GLU	C42
	THR	C43

Tableau 15 : définition des zones du domaine HTH du répresseur CRO intéragissant spécifiquement avec l'ADN

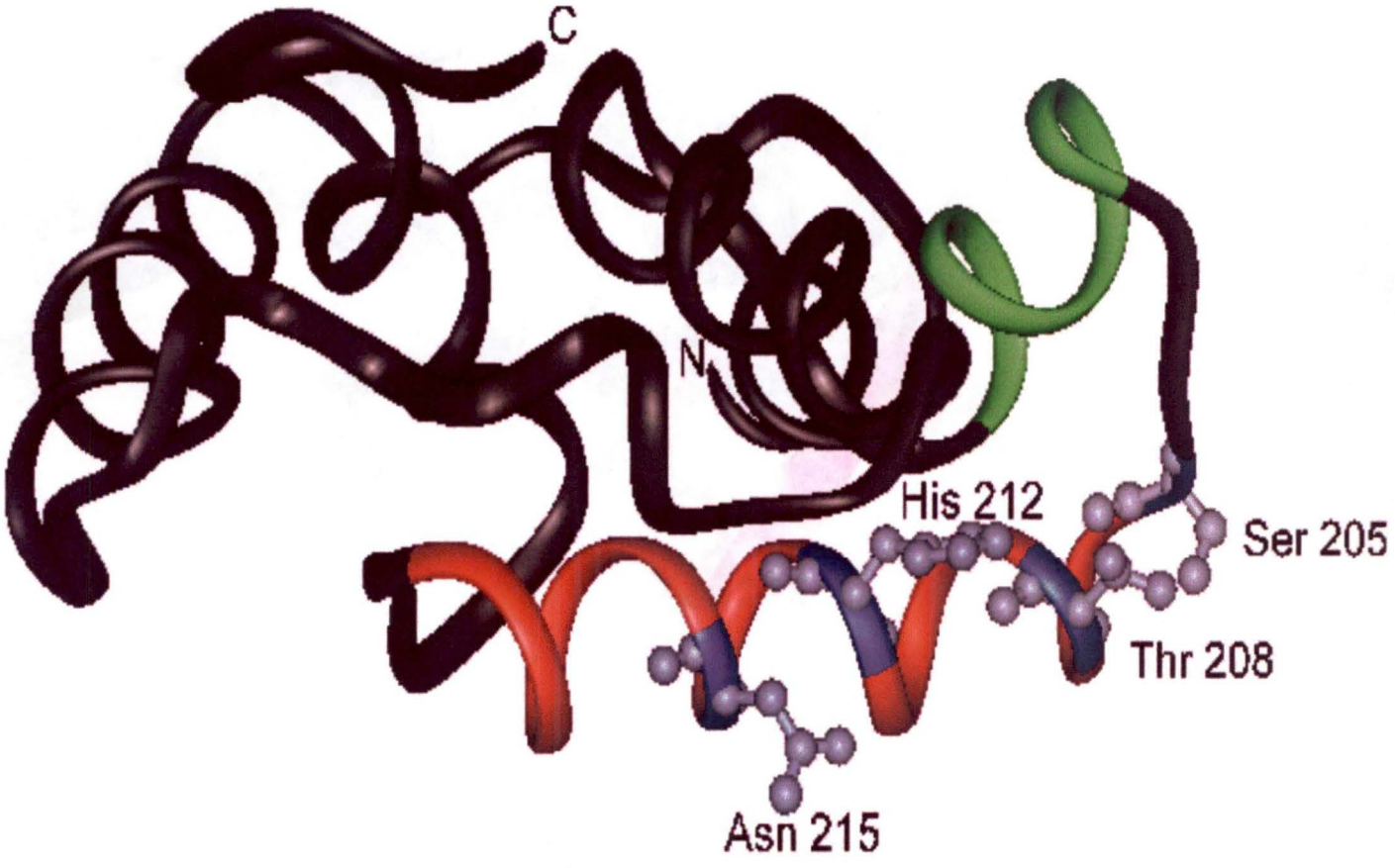


Fig. 35: Position des résidus Ser 205, Thr 208, His 212 et Asn 215 dans le modèle tridimensionnel du domaine C-terminal de BabR. La première hélice du motif HTH est colorée en vert. La seconde hélice du motif est colorée en rouge. Les résidus d'intérêt sont colorés en bleu. Les lettres N et C indiquent les extrémités N- et C-terminales du domaine.

- un répresseur lambda (code PDB du complexe : 1lmd)
- un répresseur lexa (code PDB du complexe : 1qaa)
- un répresseur P22 (code PDB du complexe : 1qar)
- un répresseur cro (code PDB du complexe : 1cro)

Nous avons analysé les interactions de ces quatre domaines HTH avec leur boîte d'ADN correspondante pour en retirer des informations extrapolables à BabR et à LuxR. Cette analyse a consisté à visualiser *via* InsightII, pour les quatre complexes, les résidus des domaines HTH formant des ponts Hydrogène avec les bases de l'ADN tout en respectant un « cut-off » de 3,5 Å maximum. Nous ne nous sommes pas intéressés aux interactions non-spécifiques c'est-à-dire celles existant entre les résidus et le squelette de l'ADN. Les résidus interagissant de manière spécifique avec l'ADN ont été consignés dans le tableau 15. Nous remarquons que ce sont toujours les mêmes types de résidus qui interviennent dans ces interactions : la glutamine, l'asparagine, le glutamate, la serine et l'histidine. De plus, comme nous pouvions nous y attendre, ce sont essentiellement les résidus de la deuxième hélice des domaines HTH qui fournissent les interactions spécifiques avec l'ADN ; en effet, c'est toujours la deuxième hélice des domaines HTH qui rentre dans le grand sillon de l'ADN.

L'analyse des complexes a permis de récolter des informations que nous pouvons maintenant mettre en relation avec les structures de BabR et LuxR. Si nous essayons de retrouver, dans la deuxième hélice α des domaines HTH de BabR et LuxR, les acides aminés appartenant au groupe des cinq types de résidus intervenant dans les interactions HTH-ADN[1], nous observons que certains d'entre eux sont conservés dans BabR et LuxR. C'est le cas de la sérine 205, de la thréonine[2] 208, de l'histidine 212 et de l'asparagine 215. Un détail intéressant est que ces résidus se trouvent environ à 3 acides aminés d'intervalle et ils sont donc placés du même côté de l'hélice α, celui qui est tourné vers l'ADN (figure 35) ; cette observation renforce notre idée que ce sont bien ces quatre résidus de BabR et LuxR qui interviendraient dans la liaison à l'ADN. D'autres résidus polaires se trouvent dans la deuxième hélice du domaine HTH, ils pourraient ainsi jouer un rôle secondaire dans la liaison à l'ADN.

La visualisation, *via* InsightII, des interactions spécifiques HTH-ADN nous a également permis de proposer les bases impliquées dans ces liens. Toutes ces protéines se lient à l'ADN sous forme d'un dimère ; il n'est ainsi pas étonnant que la boîte d'ADN soit à chaque fois un palindrôme. Nous pouvons voir que, dans trois complexes sur quatre, une thymine suivie d'une adénine se retrouvent impliquées dans la liaison au domaine HTH. De plus, la boîte d'ADN à laquelle se lierait NarL a été prédite comme étant TACYN avec Y=pyrimidine et N=amide de la chaîne principale (Baikalov *et al.*, 1996) ; la succession TA s'y retrouve. Dans chaque

[1] Ceux que nous avons déterminés en visualisant les quatre complexes HTH-ADN : la glutamine, l'asparagine, le glutamate, la serine et l'histidine.

[2] Nous pouvons la considérer aussi comme un résidu intervenant dans les interactions HTH-ADN car serine et threonine sont deux acides aminés fort proches.

lux R		
Struct. Sec.	A. aminés	numéro
	SER	41
H	SER	42
H	TRP	43
H	ASP	44
H	ILE	45
H	SER	46
H	LYS	47
H	ILE	48
H	LEU	49
T	GLY	50
T	CYS	51
T	SER	52
H	LYS	53
H	ARG	54
H	THR	55
H	VAL	56
H	THR	57
H	PHE	58
H	HIS	59
H	LEU	60
H	THR	61
H	ASN	62
H	ALA	63
H	GLN	64
H	MET	65
H	LYS	66
H	LEU	67
	ASN	68

Bab R		
Struct. Sec.	A. aminés	numéro
	THR	56
H	ALA	57
H	GLU	58
H	ILE	59
H	ILE	60
H	GLY	61
H	THR	62
H	ILE	63
H	LEU	64
T	ASN	65
T	ILE	66
T	SER	67
H	THR	68
H	ARG	69
H	THR	70
H	VAL	71
H	ASN	72
H	PHE	73
H	HIS	74
H	ILE	75
H	ASN	76
H	ASN	77
H	VAL	78
H	LEU	79
H	THR	80
H	LYS	81
H	LEU	82
	VAL	83

Tableau 16 : Prédiction des zones du domaine HTH des régulateurs LuxR et BAbR intéragissant spécifiquement avec l'ADN

Tableau 17

THREADER 2.5

domaine N-terminal

			CLASSIFICATION CATH			
	code PDB	**Z-score**	**class**	**architecture**	**topology**	**homologous superfamily (PDB code)**
1	1volA0	2.93	Mainly Alpha	Non-Bundle	Cyclin A, domain 1	1vin domain 1
2	1vhrA0	2.78	Alpha Beta	Complex	Protein-Tyrosine Phosphatase, subunit A	1ytn
3	1lam01	2.54	Alpha Beta	3-Layer(aba) Sandwich	Leucine Aminopeptidase, subunit E, domain 1	1lam domain 1
4	1fil00	2.45	Alpha Beta	2-Layer Sandwich	Beta-*Lactamase	1fil
5	1pvdA1	2.24	Alpha Beta	3-Layer(aba) Sandwich	Rossmann fold (Nitrogenase Molybdenum-Iron Protein, subunit A, domain 3)	1pox chain A domain 3

Tableau 18

3DPSSM

domaine N-terminal

			CLASSIFICATION CATH			
	code PDB	**E-value**	**class**	**architecture**	**topology**	**homologous superfamily (PDB code)**
1	1ae9 A0	1.33e-02	Mainly Alpha	Non-Bundle	Lambda Integrase; Chain: A, domain 2	1ae9 chain A domain 2
2	1pvd B1	6.84e-01	Alpha Beta	3-Layer(aba) Sandwich	hn fold (Nitrogenase Molybdenum-Iron Protein, subunit A, d	1pox chain A domain 3
3	1kln A1	1.00e+00	/	/	/	/
4	1jul _0	2.36e+00	Alpha Beta	Barrel	TIM Barrel	1pii domain 2
5	1mty G0	2.38e+00	Mainly Alpha	Bundle	Methane Monooxygenase Hydroxylase, Chain G, domain 1	1mty chain G domain 1

Tableau 19

GENTHREADER

domaine N-terminal

			CLASSIFICATION CATH			
	code PDB	probabilité	class	architecture	topology	homologous superfamily (PDB code)
1	1cot_00	0,155	Mainly Alpha	Non-Bundle	Arc Repressor Mutant, subunit A	1gks
2	1iea_A0	0,105	Alpha Beta	2-Layer Sandwich	Class II Histocompatibility Protein, subunit A, domain 1	Homologous superfamily
3	1ebd_A0	0,088	Alpha Beta	3-Layer(aba) Sandwich	Rossmann fold (Nitrogenase Molybdenum-Iron Protein, subunit A, domain 3)	3lad chain A domain 2
4	1lpf_A0	0,058	Alpha Beta	3-Layer(aba) Sandwich	Rossmann fold (Nitrogenase Molybdenum-Iron Protein, subunit A, domain 3)	3lad chain A domain 2
5	1mut_00	0,054	Alpha Beta	Complex	Nucleoside Triphosphate Pyrophosphohydrolase	1mut

Tableau 20

TOPITS

domaine N-terminal

			CLASSIFICATION CATH			
	code PDB	Z-score	class	architecture	topology	homologous superfamily (PDB code)
1	2kau_C	2.19	Alpha Beta	Barrel	Urease, subunit C, domain 2	2kau chain C domain 2
2	1tht_A	2.12	Alpha Beta	3-Layer(aba) Sandwich	Rossmann fold (Nitrogenase Molybdenum-Iron Protein, subunit A, domain 3)	1tht chain A
3	1qba	1.87	Alpha Beta	Barrel	Chitobiase, domain 3	1qba domain 3
4	2btv_A	1.87	/	/	/	/
5	1lml	1.81	Alpha Beta	Complex	Leishmanolysin , domain 2	1lml domain 2

Tableau 21

PSI-BLAST BORK

domaine N-terminal

		CLASSIFICATION CATH			
code PDB	E-value	class	architecture	topology	homologous superfamily (PDB code)
/	/	/	/	/	/

complexe, la région séparant les deux zones de liaison sur le palindrôme fait environ 10 bases de long. Puisque deux grands sillons d'ADN sont séparés par 10 bases, les deux monomères de chaque dimère interagissent avec deux grands sillons successifs. Ces données nous permettent de déterminer, sur la *lux*box, les deux régions avec lesquelles pourraient interagir spécifiquement les deux monomères du dimère LuxR-LuxR (tableau 16). Si nous connaissions la boîte de liaison de BabR, nous pourrions aussi préciser les régions d'interaction.

Nous avons pu construire un modèle du domaine C-terminal de BabR par homologie avec NarL. Des méthodes de prédictions de « fold » nous ont permis de déterminer un alignement séquence-structure optimal. Un modèle de LuxR a pu être obtenu facilement. Celui-ci ainsi qu'une analyse des interactions entre HTH et ADN de quatre complexes issus de PDB nous a permis de retirer des informations sur le domaine HTH du modèle de BabR. L'analyse des interactions HTH-ADN nous a aussi amené à proposer deux régions de la *lux*box auxquelles se lieraient le dimère LuxR-LuxR.

5. Le domaine N-terminal de BabR

Nous allons tenter ici d'obtenir le plus d'informations structurales possibles sur le domaine N-terminal de BabR. Rappelons que le domaine N-terminal de BabR lie une phéromone, fort probablement une dDHL. D'un point de vue structural, aucun domaine de ce type n'a été caractérisé auparavant. Nous avons donc commencé par rechercher, *via* BLASTP, un ou plusieurs homologues du domaine N-terminal dans la banque PDB. Aucun homologue significatif n'est ressorti. Le seul « hit » trouvé est une malate synthase dont la E-value correspondante est beaucoup trop élevée (8,1), ce qui signifie, en terme de probabilité.... Puisque ce patron n'est pas utilisable, nous nous sommes tournés vers la prédiction de structures tertiaires par d'autres méthodes, celles utilisées précédemment pour le domaine C-terminal, c'est-à-dire THREADER 2.5, 3DPSSM, GENTHREADER, TOPITS et PSI-BLAST BORK. Les résultats sont présentés dans les tableaux 17 à 21. La première observation de ces résultats ne permet pas de prédire une structure précise. Soit, elle est noyée dans le bruit de fond soit, il n'existe dans les banques de structures tertiaires aucune protéine adoptant une structure similaire. Cependant, un fait intéressant est observé : une architecture émerge des différents « hits » obtenus, qui est un motif de liaison de nucléotide ou « Rossman fold ». Cette topologie est celle adoptée, par exemple, par les « NAD-binding proteins ». Le « Rossman fold » fait partie de la classe α/β avec un plan β parallèle « twisté » ou tordu, entouré par des hélices α des deux côtés, ce qu'on appelle un sandwich à 3 couches $\alpha\beta\alpha$. Nous n'obtiendrons donc pas un modèle 3D de ce domaine de par l'absence d'une prédiction précise d'un « fold » approprié ; nous n'avons pu faire de prédictions qu'à un niveau de structure inférieur à celui

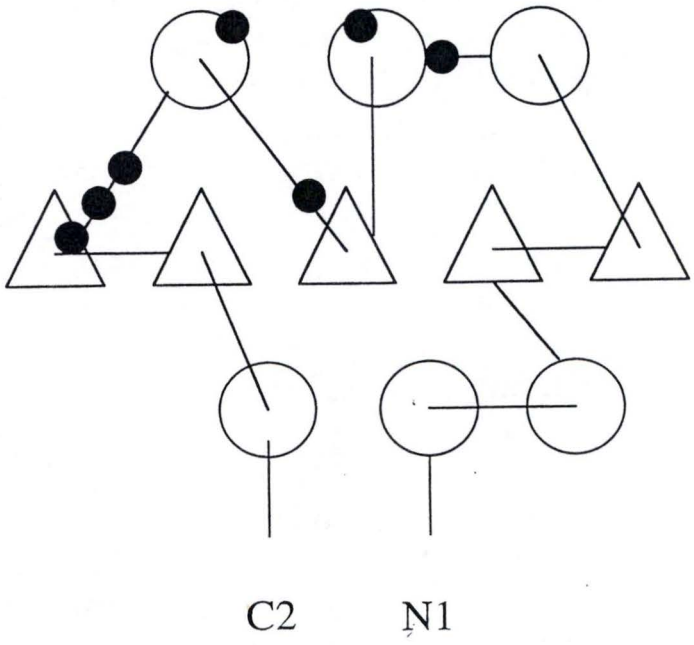


Figure 36 : La topologie prédite pour le domaine N-terminal. D'après les données de mutation que l'on possède sur LuxR et selon celles que l'on possède sur l' alignement BabR-LuxR, nous déduisons les résidus pouvant jouer un rôle important dans la liaison à l'HSL chez BabR : du domaine N-terminal au domaine C-terminal : et D76, L79, W91, I102, L111, G114 et S116

requis pour obtenir un modèle 3D. Cependant, nous essayerons de déterminer la représentation 2D éventuelle de la structure du domaine N-terminal grâce aux diagrammes de topologies fournis par le programme TOPS. Nous observerons les diagrammes TOPS des différentes protéines de type « Rossman fold », prédites par les cinq méthodes de prédictions de « folds » (cfr tableaux 17 à 21), et nous déterminerons celui qui correspond le mieux à la succession des structures secondaires prédites pour le domaine N-terminal de BabR. Par cette approche, le consensus de 2 méthodes (3DPSSM et THREADER2.5) nous a permis de déterminer que le domaine 1 de la pyruvate décarboxylase (code PDB :1pvd) est celui qui correspond le mieux au domaine N-terminal de BabR. Les résidus du domaine N-terminal de LuxR qui ont prouvé leur importance dans la liaison à l'HSL par des expériences de mutagenèse dirigée et qui, en plus, se retrouvent conservés dans le domaine N-terminal de BabR ont été placés sur la figure 36. Leur position sur le modèle topologique montrerait le site de liaison à l'HSL de BabR.

Nous n'avons finalement pas pu produire de modèle tridimensionnel du domaine N-terminal de BabR par manque d'un patron adéquat. Nous avons néanmoins pu fournir quelques caractéristiques structurales intéressantes : le domaine N-terminal adopterait la même topologie que les « NAD-binding proteins » et la région éventuelle de liaison à l'HSL a été localisée.

CONCLUSIONS ET PERSPECTIVES

Notre travail a consisté en la caractérisation structurale du régulateur transcriptionnel BabR de *Brucella abortus*. Ce régulateur fait partie de la famille LuxR dont les membres sont impliqués dans un phénomène dépendant de la densité cellulaire, le « Quorum Sensing ». Une phéromone nommée N-acyl-L-homoserine lactone (HSL ou autoinducteur) est produite par les cellules bactériennes à un taux basal. Quand la densité cellulaire dépasse un certain seuil, la phéromone atteint une concentration suffisante pour activer le régulateur transcriptionnel. Le plus étudié à ce jour est LuxR de *Vibrio fischeri*. Il peut être scindé en deux domaines, le domaine N-terminal impliqué dans la liaison à l'HSL et le domaine C-terminal impliqué dans la liaison à l'ADN.

Au terme de ce travail, nous avons établi un modèle tridimensionnel du domaine C-terminal de BabR sur base de l'homologie avec NarL, une protéine régulatrice de la transcription intervenant dans un système à deux composantes chez *E. coli*. Une fois ce modèle affiné par minimisation d'énergie potentielle, une analyse des interactions HTH-ADN de différents complexes a permis de tirer plusieurs conclusions sur le domaine HTH de BabR. Celui-ci a été assigné aux hélices α 9 et 10 grâce au modèle.

Les nombreuses informations disponibles pour LuxR ont été d'une aide capitale pour l'étude de BabR :

- nous avons ainsi pu proposer les résidus du domaine HTH de BabR qui lieraient l'ADN de manière spécifique : la sérine 205, la thréonine 208, l'histidine 212 et l'asparagine 215.
- quelques pistes concernant les boîtes de type *lux*box ont également pu être tirées de cette analyse : localisation des régions de la *lux*box auxquelles se fixerait de manière spécifique le dimère LuxR-LuxR ; le motif TA se retrouve presqu'à chaque fois dans les régions de liaison spécifique à un domaine HTH.

En ce qui concerne le domaine N-terminal de BabR, la stratégie de modélisation par homologie n'a pu être appliquée en raison d'un manque d'homologues de structure tridimensionnelle connue. A ce jour, aucun domaine de ce type n'a été décrit d'un point de vue structural. Nous nous sommes donc rabattus sur les méthodes de prédictions de « folds ».

- bien qu'aucun « fold » ne soit clairement ressorti des résultats, une topologie semblait quand même prédominer, le « **Rossman fold** ». Cette topologie est adoptée, entre autres, par les protéines liant le NAD. Dans le cas de BabR, c'est une phéromone de type homoserine lactone qui est liée par la protéine.

- nous avons finalement pu obtenir une représentation 2D de la structure qu'adopterait le domaine N-terminal de BabR grâce à une analyse des diagrammes TOPS des différents « Rossman fold » prédits. Le « Rossman fold » qui nous a semblé le plus probable est 1pvd, une **pyruvate decarboxylase**.
- les données de mutation que l'on possède sur LuxR et l'alignement LuxR-BabR issu d'un alignement multiple, nous ont permis de localiser, d'un point de vue structural, la région à laquelle se lierait la phéromone. Nous avons donc une hypothèse concernant la structure 3D et la fonction de ce domaine, ce qui n'avait encore jamais été fait.

Toutes ces prédictions doivent évidemment être validées expérimentalement. Par exemple, pour le domaine C-terminal de BabR, les résidus S205, T208, H212 et N215 de BabR pourront être testés par des expériences de mutagenèse dirigée. Nous pourrions aussi muter les bases de la région de la *lux*box pouvant intervenir dans la liaison spécifique au HTH, pour vérifier leur importance. Une analyse de « docking » protéine-ADN serait très intéressante pour récolter des données précises sur les interactions entre les protéines de la famille LuxR et les boîtes de type *lux*box ; il faudrait pour cela, la structure résolue de la protéine ou éventuellement un modèle auquel nous pourrions accorder une très bonne confiance, la boîte d'ADN à laquelle se lie la protéine et le positionnement approximatif de la protéine par rapport à l'ADN. MONTY, un programme de « docking » protéine-ADN, prend en compte le fait que la liaison de la protéine à l'ADN entraîne des changements conformationnels dans l'ADN ; ceci est appelé le « bending » de l'ADN. Si NarL était cristallisée dans un futur proche sous forme d'un complexe protéine-ADN, ce complexe pourrait servir de « template » pour positionner le modèle de LuxR par rapport à sa boîte de liaison. De plus, si la boîte de liaison préférentielle de BabR est élucidée, cette analyse de « docking » pourrait être appliquée à BabR.

Pour le domaine N-terminal de BabR, des tests de mutagenèse dirigée sur les résidus D76, L79, W91, I102, L111, G114 et S116 pourront être envisagés pour démontrer leur importance dans la liaison à l'HSL.

Cette approche de modélisation nous a amené à travailler sur les deux domaines de BabR pris séparément. Les données récoltées sur chacun des domaines individuels, bien qu'elles soient intéressantes, ne nous permettent pas de connaître les interactions éventuelles qu'il pourrait y avoir entre les deux domaines. En effet, il est probable que la liaison de l'HSL au domaine N-terminal engendre une modification des interactions avec le domaine C-terminal qui modifierait la capacité de BabR à interagir avec un promoteur cible.

BIBLIOGRAPHIE

Adar, Y.Y., Simaan, M. and Ulitzur, S. (1992) Formation of the LuxR protein in the Vibrio fischeri lux system is controlled by HtpR through the GroESL proteins. *J Bacteriol*, **174**, 7138-7143.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol*, **215**, 403-410.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25**, 3389-3402.

Attwood, T.K. and Parry-Smith, D.J. (1999) *Introduction to bioinformatics*. Addison Wesley Longman, Essex.

Baikalov, I., Schroder, I., Kaczor-Grzeskowiak, M., Cascio, D., Gunsalus, R.P. and Dickerson, R.E. (1998) NarL dimerization? Suggestive evidence from a new crystal form. *Biochemistry*, **37**, 3665-3676.

Baikalov, I., Schröder, I., Kaczor-Grzeskowiak, M., Grzeskowiak, K., Gunsalus, R.P. and Dickerson, R.E. (1996) Structure of the Escherichia coli Response Regulator NarL. *Biochemistry*, **35**, 11053-11061.

Baxevanis, A.D. and Ouellette, B.F.F. (1998) *Bioinformatics : A Practical Guide to the Analysis of Genes and Proteins*. Wiley-Interscience, New-York.

Belas, R. (1997) Proteus mirabilis and other swarming bacteria. In Shapiro, J.A. and Dworkin, M. (eds.), *Bacteria as Multicellular Organisms*. Oxford University Press, Oxford, New-York, pp. 183-219.

Ben-Jacob, E. and Cohen, I. (1997) Cooperative formation of bacterial patterns. In Shapiro, J.A. and Dworkin, M. (eds.), *Bacteria as Multicellular Organisms*. Oxford University Press, Oxford, New-York, pp. 394-416.

Beppu, T. (1995) Signal transduction and secondary metabolism: prospects for controlling productivity. *Trends Biotechnol*, **13**, 264-269.

Branden, C. and Tooze, J. (1991) *Introduction to Protein Structure*. Garland Publishing, New-York.

Brasseur, R. (1995) Simulating the folding of small proteins by use of the local minimum energy and the free solvation energy yields native-like structures. *J Mol Graph*, **13**, 312-322.

Budrene, E.O. and Berg, H.C. (1995) Dynamics of formation of symmetrical patterns by chemotactic bacteria. *Nature*, **376**, 49-53.

Busby, S. and Ebright, R.H. (1997) Transcription activation at class II CAP-dependent promoters. *Mol Microbiol*, **23**, 853-859.

Chater, K.F. and Losick, R. (1997) The mycelial life-style of Streptomyces coelicolor A3(2) and its relatives. In Shapiro, J.A. and Dworkin, M. (eds.), *Bacteria as Multicellular Organisms*. Oxford University Press, Oxford, New-York, pp. 149-182.

Choi, S.H. and Greenberg, E.P. (1992a) Genetic dissection of DNA binding and luminescence gene activation by the Vibrio fischeri LuxR protein. *J Bacteriol*, **174**, 4064-4069.

Choi, S.H. and Greenberg, E.P. (1992b) Genetic evidence for multimerization of LuxR, the transcriptional activator of Vibrio fischeri luminescence. *Mol. Mar. Biol. Biotechnol.*, **1**, 408-413.

Chou, P.Y. and Fasman, G.D. (1974) Prediction of protein conformation. *Biochemistry*, **13**, 211-222.

Clewell, D.B. (1993) Bacterial sex pheromone-induced plasmid transfer. *Cell*, **73**, 9-12.

Costerton, J.W., Lewandowski, Z., Caldwell, D.E., Korber, D.R. and Lappin-Scott, H.M. (1995) Microbial biofilms. *Annu Rev Microbiol*, **49**, 711-745.

Cuff, J.A., Clamp, M.E., Siddiqui, A.S., Finlay, M. and Barton, G.J. (1998) JPred: a consensus secondary structure prediction server. *Bioinformatics*, **14**, 892-893.

Davies, D.G., Parsek, M.R., Pearson, J.P., Iglewski, B.H., Costerton, J.W. and Greenberg, E.P. (1998) The involvement of cell-to-cell signals in the development of a bacterial biofilm [see comments]. *Science*, **280**, 295-298.

Dayhoff, M.O., E.R.V. and Park, C.M. (1972) A model of evolutionary change in proteins. *Atlas of protein sequence and structure*, **5**, 88-99.

de Fays, K., Tibor, A., Lambert, C., Vinals, C., Denoël, P., De Bolle, X., Wouters, J., Letesson, J.-J. and Depiereux, E. (1999) Structure and function prediction of the *Brucella abortus* **P39** protein by comparative modeling with marginal sequence similarities. *Protein Engineering*, **12**, 217-223.

Depiereux, E., Baudoux, G., Briffeuil, P., Reginster, I., De Bolle, X., Vinals, C. and Feytmans, E. (1997) Match-Box_server: a multiple sequence alignment tool placing emphasis on reliability. *Comput Appl Biosci*, **13**, 249-256.

Depiereux, E. and Feytmans, E. (1992) MATCH-BOX: a fundamentally new algorithm for the simultaneous alignment of several protein sequences. *Comput Appl Biosci*, **8**, 501-509.

Devine, J.H., Shadel, G.S. and Baldwin, T.O. (1989) Identification of the operator of the lux regulon from the Vibrio fischeri strain ATCC7744. *Proc Natl Acad Sci U S A*, **86**, 5688-5692.

DISCOVER User Guide, San Diego : MSI (1995a)

Doolittle, R.F. (1981) Similar amino acid sequences: chance or common ancestry? *Science*, **214**, 149-159.

Dunlap, P.V. (1997) N-Acyl-l-Homoserine Lactone Autoinducers in Bacteria. In Shapiro, J.A. and Dworkin, M. (eds.), *Bacteria as Multicellular Organisms*. Oxford University Press, Oxford, New-York, pp. 69-106.

Dunny, G.M. and Winans, S.C. (1999) Bacterial life: neither lonely nor boring. In Dunny, G.M. and Winans, S.C. (eds.), *Cell-Cell Signaling in Bacteria*. ASM Press, Washington, D.C., pp. 1-5.

Dworkin, M. (1997) Multiculturism Versus the Single Microbe. In Shapiro, J.A. and Dworkin, M. (eds.), *Bacteria as Multicellular Organisms*. Oxford University Press, Oxford, New-York, pp. 3-13.

Eberhard, A., Burlingame, A.L., Eberhard, C., Kenyon, G.L., Nealson, K.H. and Oppenheimer, N.J. (1981) Structural identification of autoinducer of Photobacterium fischeri luciferase. *Biochemistry*, **20**, 2444-2449.

Eberhard, A., Longin, T., Widrig, C.A. and Stranick, S.J. (1991) Synthesis of the lux gene autoinducer in Vibrio fischeri is positively autoregulated. *Arch. Microbiol.*, **155**, 294-297.

Engebrecht, J., Nealson, K. and Silverman, M. (1983) Bacterial bioluminescence: isolation and genetic analysis of functions from Vibrio fischeri. *Cell*, **32**, 773-781.

Engebrecht, J. and Silverman, M. (1984) Identification of genes and gene products necessary for bacterial bioluminescence. *Proc Natl Acad Sci U S A*, **81**, 4154-4158.

Fischer, D., Barret, C., Bryson, K., Elofsson, A., Godzik, A., Jones, D., Karplus, K.J., Kelley, L.A., MacCallum, R.M., Pawowski, K., Rost, B., Rychlewski, L. and Sternberg, M. (1999) CAFASP-1: critical assessment of fully automated structure prediction methods. *Proteins*, **Suppl**, 209-217.

Fuqua, C. and Eberhard, A. (1999) Signal generation in Autoinduction systems: synthesis of acylated homoserine lactones by LuxI-type proteins. In Dunny, G.M. and Winans, S.C. (eds.), *Cell-Cell Signaling in Bacteria*. ASM Press, Washington, D.C., pp. 211-230.

Fuqua, W.C., Winans, S.C. and Greenberg, E.P. (1994) Quorum sensing in bacteria: the LuxR-LuxI family of cell density-responsive transcriptional regulators. *Journal of Bacteriology*, **176**, 269-275.

Gambello, M.J. and Iglewski, B.H. (1991) Cloning and characterization of the Pseudomonas aeruginosa lasR gene, a transcriptional activator of elastase expression. *J Bacteriol*, **173**, 3000-3009.

Garnier, J., Osguthorpe, D.J. and Robson, B. (1978) Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J Mol Biol*, **120**, 97-120.

Gilis, D. (1999) Prédiction du repliement et de la stabilité de protéines par des potentiels dérivés de structures connues. *Thèse de Doctorat, ULB*.

Gilis, D. and Rooman, M. (2000) Identification and ab initio simulations of early folding units in proteins. *submitted.*

Givskov, M., Östling, J., Ebert, L., Lindum, P.W. and Christensen, A.B. (1998) The participation of two separate regulatory systems in controling swarming motility of Serratia liquefaciens. *J. Bacteriol.*, **180**.

Goodsell, D.S. and Olson, A.J. (1990) Automated Docking of Substrates to Proteins by Simulated Annealing. *Proteins*, **8**, 195-202.s

Greenberg, E.P. (1997) Quorum Sensing in Gram-Negative Bacteria. *ASM News*, **63**, 371-377.

Grossman, A.D. (1995) Genetic networks controlling the initiation of sporulation and the development of genetic competence in Bacillus subtilis. *Annu Rev Genet*, **29**, 477-508.

Guex, N. and Peitsch, M.C. (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis*, **18**, 2714-2723.

Gygi, D., Rahman, M.M., Lai, H.C., Carlson, R., Guard-Petter, J. and Hughes, C. (1995) A cell-surface polysaccharide that facilitates rapid population migration by differentiated swarm cells of Proteus mirabilis. *Mol Microbiol*, **17**, 1167-1175.

Hanzelka, B.L., Stevens, A.M., Parsek, M.R., Crone, T.J. and Greenberg, E.P. (1997) Mutational analysis of the Vibrio fischeri LuxI polypeptide: critical regions of an autoinducer synthase. *J Bacteriol*, **179**, 4882-4887.

Henikoff, S. and Henikoff, J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, **89**, 10915-10919.

Hofmann, K. and Stoffel, W. (1993) TMbase - A database of membrane spanning proteins segments. *Biol. Chem. Hoppe-Seyler*, **347**.

Holm, L. and Sander, C. (1992) Fast and simple Monte Carlo algorithm for side chain optimization in proteins: application to model building by homology. *Proteins*, **14**, 213-223.

Hubbard, T., Park, J., Lahm, A., Leplae, R. and Tramontano, A. (1996) Protein structure prediction: playing the fold. *Trends Biochem Sci*, **21**, 279-281.

Huynen, M., Doerks, T., Eisenhaber, F., Orengo, C., Sunyaev, S., Yuan, Y. and Bork, P. (1998) Homology-based fold predictions for Mycoplasma genitalium proteins. *J Mol Biol*, **280**, 323-326.

Ishihama, A. (1993) Protein-protein communication within the transcription apparatus. *J Bacteriol*, **175**, 2483-2489.

InsightII User Guide. , San Diego : MSI. (1995b)

Jones, D.T. (1999a) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J Mol Biol*, **287**, 797-815.

Jones, D.T. (1999b) Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol*, **292**, 195-202.

Jones, D.T., Tress, M., Bryson, K. and Hadley, C. (1999) Successful recognition of protein folds using threading methods biased by sequence similarity and predicted secondary structure. *Proteins*, **Suppl**, 104-111.

Jones, S., Yu, B., Bainton, N.J., Birdsall, M., Bycroft, B.W., Chhabra, S.R., Cox, A.J., Golby, P., Reeves, P.J., Stephens, S. and et al. (1993) The lux autoinducer regulates the production of exoenzyme virulence determinants in Erwinia carotovora and Pseudomonas aeruginosa. *Embo J*, **12**, 2477-2482.

Kaplan, H.B. and Greenberg, E.P. (1987) Overproduction and purification of the luxR gene product : transcriptional activator of the Vibrio fischeri luminescence system. *Proc. Natl. Acad. Sci.*, **84**, 6639-6643.

Kaplan, H.B. and Plamann, L. (1996) A Myxococcus xanthus cell density-sensing system required for multicellular development. *FEMS Microbiol Lett*, **139**, 89-95.

Kelley, L.A., MacCallum, R.M. and Sternberg, M.J.E. (2000) Enhanced Genome Annotation using Structural Profiles in the Program 3DPSSM. *J. Mol. Biol.*, **299**, 501-522.

King, R.D., Ouali, M., Strong, A.T., Aly, A., Elmaghraby, A., Kantardzic, M. and Page, D. (2000) Is it better to combine predictions ? *Protein Engineering*, **13**, 15-19.

King, R.D., Saqi, M., Sayle, R. and Sternberg, M.J. (1997) DSC: public domain protein secondary structure predication. *Comput Appl Biosci*, **13**, 473-474.

King, R.D. and Sternberg, M.J. (1990) Machine learning approach for the prediction of protein secondary structure. *J Mol Biol*, **216**, 441-457.

Kleerebezem, M., Quadri, L.E. and Kuipers de vos, O.P. (1997) Quorum sensing by peptide pheromones and two component signal-transduction systems in Gram-positive bacteria. *Mol.Microbiol.*, **24**, 895-904.

Knegtel, R.M.A., Antoon, J., Rullmann, C., Boelens, R. and Kaptein, R. (1994) MONTY: a Monte Carlo Approach to Protein-DNA Recognition. *J.Mol.Biol.*, **235**, 318-324.

Kolibachuk, D. and Greenberg, E.P. (1993) The Vibrio fischeri luminescence gene activator LuxR is a membrane-associated protein. *J Bacteriol*, **175**, 7307-7312.

Leach, A.R. (1994) Ligand docking to proteins with discrete side-chain flexibility. *J Mol Biol*, **235**, 345-356.

Lengauer, T. and Rarey, M. (1996) Computational methods for biomolecular docking. *Curr Opin Struct Biol*, **6**, 402-406.

Lim, V.I. (1974) Structural principles of the globular organization of proteins chains. A stereochemical theory of globular proteins secondary structure. *J. Mol. Biol.*, **88**, 857-872.

Losick, R. and Kaiser, D. (1997) La communication des bactéries. *Pour la science*, **234**, 76-82.

Luthy, R., Bowie, J.U. and Eisenberg, D. (1992) Assessment of protein models with three-dimensional profiles. *Nature*, **356**, 83-85.

McGowan, S., Sebaihia, M., Jones, S., Yu, B., Bainton, N., Chan, P.F., Bycroft, B., Stewart, G.S., Williams, P. and Salmond, G.P. (1995) Carbapenem antibiotic production in Erwinia carotovora is regulated by CarR, a homologue of the LuxR transcriptional activator [published erratum appears in Microbiology 1995 May;141(Pt 5):1268]. *Microbiology*, **141**, 541-550.

Miller, R.T., Jones, D.T. and Thornton, J.M. (1996) Protein fold recognition by sequence threading: tools and assessment techniques. *Faseb J*, **10**, 171-178.

Mizutani, M.Y., Tomioka, N. and Itai, A. (1994) Rational automatic search method for stable docking models of protein and ligand. *J Mol Biol*, **243**, 310-326.

Moré, M.I., Finger, L.D., Stryker, J.L., Fuqua, C., Eberhard, A. and Winans, S.C. (1996) Enzymatic synthesis of a quorum-sensing autoinducer through use of defined substrates. *Science*, **272**, 1655-1658.

Morris, G.M., Goodsell, D.S., Halliday, R.S., Huey, R., Hart, W.E., Belew, R.K. and Olson, A.J. (1998) Automated docking using Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comp. Chem.*, **19**, 1639-1662.

Muggleton, S., King, R.D. and Sternberg, M.J. (1992) Protein secondary structure prediction using logic-based machine learning [published erratum appears in Protein Eng 1993 Jul;6(5):549]. *Protein Eng*, **5**, 647-657.

Myers, E.W. and Miller, W. (1988) Optimal alignments in linear space. *Comput Appl Biosci*, **4**, 11-17.

Nakai, K. and Horton, P. (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem Sci*, **24**, 34-36.

Nakai, K. and Kanehisa, M. (1991) Expert system for predicting protein localization sites in Gram-Negative bacteria. *PROTEINS*, **11**, 95-110.

Olmea, O. and Valencia, A. (1997) Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Fold Des*, **2**, S25-32.

Ouali, M. and King, R.D. Cascaded Multiple Classifers for Secondary Structure Prediction. *Submitted for publication.*

Overington, J., Johnson, M.S., Sali, A. and Blundell, T.L. (1990) Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. *Proc R Soc Lond B Biol Sci*, **241**, 132-145.

Pearson, W.R. and Lipman, D.J. (1988) Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A*, **85**, 2444-2448.

Pierson III, L.S., Wood, D.W. and von Bodman, S.B. (1999) Quorum Sensing in plant-associated bacteria. In Dunny, G.M. and Winans, S.C. (eds.), *Cell-Cell Signaling in Bacteria*. ASM Press, Washington, D.C., pp. 101-115.

Puskas, A., Greenberg, E.P., Kaplan, S. and Schaefer, A.L. (1997) A quorum-sensing system in the free-living photosynthetic bacterium Rhodobacter sphaeroides. *J Bacteriol*, **179**, 7530-7537.

Rost, B. (1995) TOPITS: threading one-dimensional predictions into three-dimensional structures. *Ismb*, **3**, 314-321.

Rost, B. and Sander, C. (1993) Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol*, **232**, 584-599.

Russel, R.B. (1999) A guide to structure prediction (Version 2). *http://www.bmm.icnet.uk/people/rob/CCP11BBS/index.html*.

Sanchez, R. and Sali, A. (1997) Evaluation of comparative protein structure modeling by MODELLER-3. *Proteins*, **Suppl**, 50-58.

Schaefer, A.L., Hanzelka, B.L., Eberhard, A. and Greenberg, E.P. (1996) Quorum sensing in Vibrio fischeri: probing autoinducer-LuxR interactions with autoinducer analogs. *J Bacteriol*, **178**, 2897-2901.

Shadel, G.S. and Baldwin, T.O. (1991) The Vibrio fischeri LuxR protein is capable of bidirectional stimulation of transcription and both positive and negative regulation of the luxR gene. *J Bacteriol*, **173**, 568-574.

Shadel, G.S., Young, R. and Baldwin, T.O. (1990) Use of regulated cell lysis in a lethal genetic selection in Escherichia coli: identification of the autoinducer-binding region of the LuxR protein from Vibrio fischeri ATCC 7744. *J Bacteriol*, **172**, 3980-3987.

Shapiro, J.A. (1998) Thinking about bacterial populations as multicellular organisms. In Ornston, L.N., Balones, A. and Greenberg, E.P. (eds.), *Annual Review of Microbiology*. Annual Reviews, Palo Alto, pp. 81-104.

Shoichet, B.K. and Kuntz, I.D. (1991) Protein docking and complementarity. *J Mol Biol*, **221**, 327-346.

Sippl, M.J. (1993) Boltzmann's Principle, Knowledge Based Mean Fields and Protein Folding. An Approach to the Computational Determination of Protein Structures. *J. Computer Aided Mol. Design*, **7**, 473-501.

Sitnikov, D.M., Shadel, G.S. and Baldwin, T.O. (1996) Autoinducer-independent mutants of the LuxR transcriptional activator exhibit differential effects on the two lux promoters of Vibrio fischeri. *Mol Gen Genet*, **252**, 622-625.

Solomon, J.M. and Grossman, A.D. (1996) Who's competent and when: regulation of natural genetic competence in bacteria. *Trends Genet*, **12**, 150-155.

Sonea, S. and Panisset, M. (1980) *Introduction à la nouvelle bactériologie*. Les Presses de l'Université de Montreal, Masson, Montreal, Paris.

Sternberg, M.J., Gabb, H.A. and Jackson, R.M. (1998) Predictive docking of protein-protein and protein-DNA complexes. *Curr Opin Struct Biol*, **8**, 250-256.

Sternberg, M.J.E. (1996) *Protein Structure Prediction*
A Practical Approach. Oxford University Press, Oxford.

Stevens, A.M., Dolan, K.M. and Greenberg, E.P. (1994) Synergistic binding of the Vibrio fischeri LuxR transcriptional activator domain and RNA polymerase to the lux promoter region [published erratum appears in Proc Natl Acad Sci U S A 1995 Apr 11;92(8):3631]. *Proc Natl Acad Sci U S A*, **91**, 12619-12623.

Stevens, A.M. and Greenberg, E.P. (1999) Transcriptional activation by LuxR. In Dunny, G.M. and Winans, S.C. (eds.), *Cell-Cell Signaling in Bacteria*. ASM Press, Washington, D.C., pp. 231-242.

Swift, S., Bainton, N.J. and Winson, M.K. (1994) Gram-negative bacterial communication by N-acyl homoserine lactones: a universal language? *Trends in Microbiology*, **2**, 193-198.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, **22**, 4673-4680.

Totrov, M. and Abagyan, R. (1994) Detailed ab initio prediction of lysozyme-antibody complex with 1.6 A accuracy. *Nat Struct Biol*, **1**, 259-263.

Ulitzer, S. and Dunlap, P.V. (1995) Regulatory circuitry controlling luminescence autoinduction in Vibrio fischeri. *Photochem. Photobiol.*, **62**, 625-632.

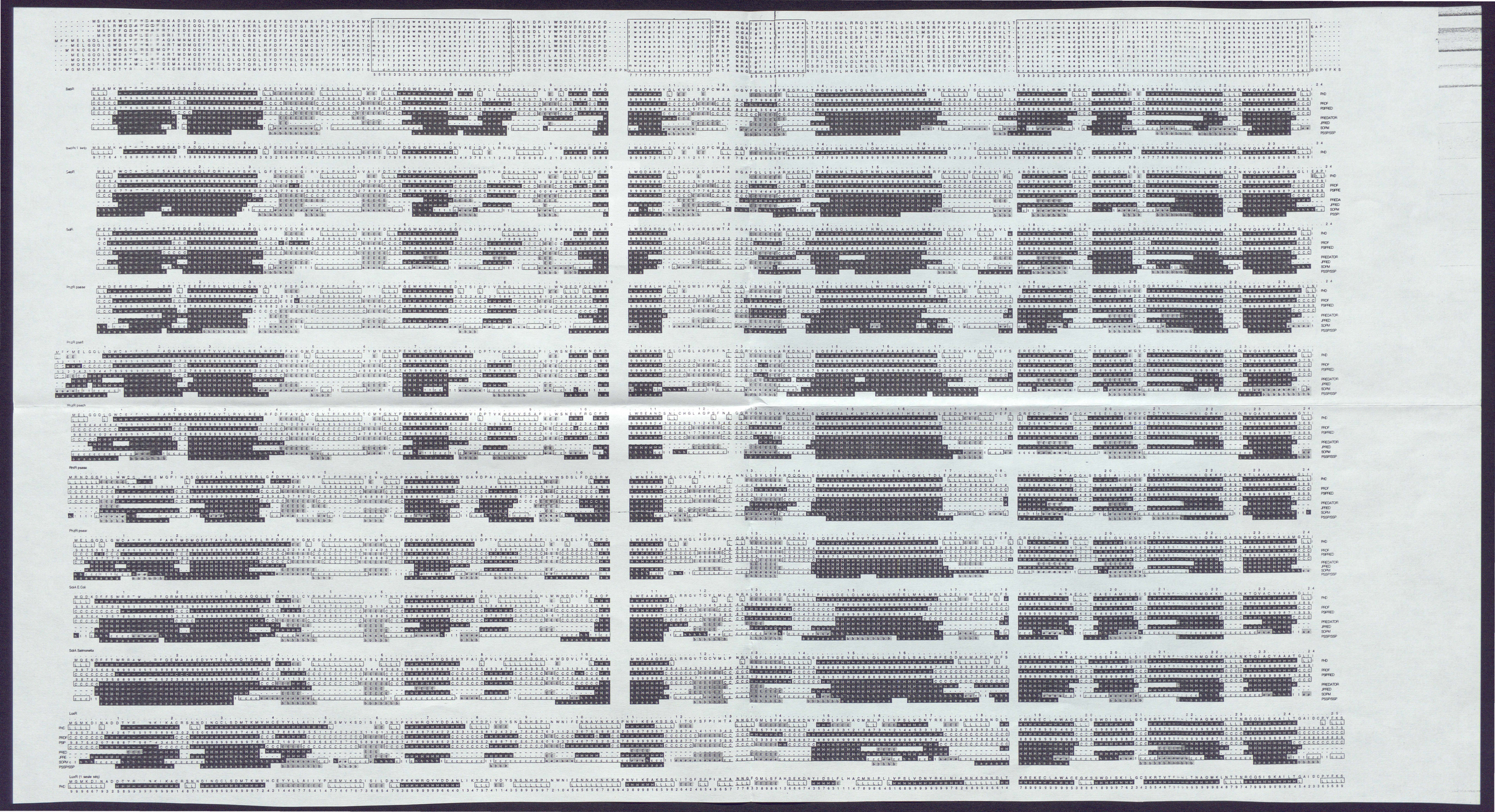
Vinals, C. (1996) Modélisation de structures tridimensionnelles de protéines : application à l'étude de la stéréospécificité des lactate déshydrogénases. . Facultés Universitaires Notre-Dame de la Paix, Namur.

Welch, M., Todd, D.E., Whitehead, N.A., McGowan, S.J., Bycroft, B.W. and Salmond, G.P.C. (2000) N-acyl homoserine lactone binding to the CarR receptor determines quorum-sensing specificity in Erwinia. *The EMBO Journal*, **19**, 631-641.

Westhead, D.R., Slidel, T.W., Flores, T.P. and Thornton, J.M. (1999) Protein structural topology: Automated analysis and diagrammatic representation. *Protein Sci*, **8**, 897-904.

Zhang, L., Murphy, P.J., Kerr, A. and Tate, M.E. (1993) Agrobacterium conjugation and gene regulation by N-acyl-L-homoserine lactones. *Nature*, **362**, 446-448.

Zhu, J. and Winans, S.C. (1999) Autoinducer binding by the quorum-sensing regulator TraR increases affinity for target promoters in vitro and decreases TraR turnover rates in whole cells. *Proc Natl Acad Sci U S A*, **96**, 4832-4837.


BatR
tsaR(1 seq)
SeqR
SqR
PmzR pseae
PmzR psefl
PmzR psech
RhtR psese
PhzR psear
SdiA E.Coli
SdiA Salmonella
LuxR
LuxR (1 seule séq)
PHD
PROF
PSIPRED
PREDATOR
JPRED
SCRATCH
PSSPSSP